This is a previous version of the article published in Applied Clay Science. 2024, 249: 107239. https://doi.org/10.1016/j.clay.2023.107239

A SUPER-LEARNER MACHINE LEARNING MODEL FOR A GLOBAL PREDICTION OF COMPRESSION INDEX IN CLAYS

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, spagnoli_giovanni@yahoo.de, giovanni.spagnoli@dmt-group.com; ORCID: 0000-0002-1866-4345

* Corresponding author.

Abstract

Settlement of structures is determined by the stiffness of the soil where they are built. Compression index (c_c) quantifies the compressibility of the soil and is a key parameter in the design of geotechnical structures. To predict the value of c_c in clay soils, a global database of more than 1000 data points was collected and analysed. Liquid limit, plasticity index, natural water content, and initial void ratio were considered as predictor variables. A super-learner machine learning model was developed to predict c_c from these variables. The model demonstrated a reasonable predictive performance and was subsequently integrated into an online tool. Additionally, four symbolic regression expressions were obtained to relate c_c with some of the input variables when not all data are available, providing simple and practical alternatives for c_c, estimation. This study provided two major contributions: (1) the non-local nature of the data expands the scope and generalizability of the findings, and (2) the availability of the proposed algorithm through an online application ensures its accessibility for geotechnical engineers, enhancing the work's practical applicability and intrinsic value.

Keywords: machine learning; compression index; liquid limit; plasticity index; natural water content; initial void ratio; clay.

1. Introduction

Compression index (c_c), which is the slope of the linear section of the e–logσ' plot (and is dimensionless), and the coefficient of consolidation (c_v), primarily define the compressibility properties of fine-grained soils (Craig 2004). c_c is a crucial parameter for characterizing soil compressibility and deformation behaviour under load. It is commonly used in geotechnical engineering to estimate the settlement and deformation of soil structures, such as foundations and embankments (Das 2021). Soils with a higher c_c tend to be more compressible and deformable under load, whereas soils with a lower c_c are less compressible and have a greater capacity to withstand deformation when they are loaded. The c_c value in soils is obtained through the oedometer test, which is relatively time-consuming and results in higher costs when compared to standard index tests. Since Atterberg limits initially characterize soils, these state parameters have been used to establish correlations with various other engineering properties of soils. Furthermore, plasticity is influenced by the electrochemical behaviour of clay minerals (Carter and Bentley 1991) in the same way as c_c (Onyejekwe et al. 2015). For this reason, several attempts were made in the past to correlate basic geotechnical properties with c_c. Numerous authors provided linear equations relating c_c to the liquid limit (LL) of soils (e.g. Azzouz et al. 1976; Bowles 1979; Park and Lee 2011; Sridharan and Nagaraj 2000; Terzaghi et al. 1967; Tsuchida 1991). On the other hand, plasticity index (PI), was also correlated with c_c (Sridharan and Nagaraj 2000; Wroth and Wood 1978). Additionally, many correlations based on a linear relationship with natural water content (w) were proposed (e.g. Azzouz et al. 1976; Koppula 1981; Rendon-Herrero 1980). The relationship with initial void ratio (e_0) was examined by Nishida (1956), Hough (1957) and (Bowles 1979), among others. Other studies included more than one index property in the estimation of c_c, such as w and LL (Koppula 1981) or e_0 and LL (Al-Khafaji and Andersland 1992). However, when these correlations were tested with new data, they exhibited significant scatter, with deviations reaching up to 30% (Spagnoli and Shimobe 2020), suggesting a lack of universally applicable validity. They are applicable within specific limits and should be restricted to the soil type or location where they were validated (Verbrugge and Schroeder 2018). Using

these correlations in different conditions may lead to unsatisfactory outcomes (Onyejekwe et al. 2016). To address the limits of classic regression approaches in geotechnical engineering, the application of machine learning (ML) algorithms have been extensively developed, demonstrating improved performance for predicting several soil engineering properties compared to traditional statistical methods (e.g. Bardhan et al. 2023; Bardhan et al. 2021; Dam Nguyen et al. 2022; Díaz and Tomás 2021; Salvatore et al. 2022; Singh et al. 2023; Trong et al. 2021). However, it is important to be aware of the limitations and uncertainties associated with ML approaches before applying them to real-world geotechnical engineering projects. Numerous studies (Baghbani et al. 2022; Zhang et al. 2023; Zhang et al. 2022) have exposed these limitations, which are primarily: a) the scarcity of high-quality data, b) the difficulty in interpreting the models, and c) the lack of generalization. Regarding data availability, geotechnical data can be costly and often incomplete or uncertain. This can lead to ML models that are not as accurate or reliable as desired. Another limitation of ML is the interpretability of the models (i.e., black box). ML models are often complex and nonlinear, making it challenging to understand the relationship between input data and output predictions. Finally, due to the inherent heterogeneity and spatial variability of soil deposits, it is difficult for empirical models trained on limited datasets to reliably extrapolate beyond the geographic scope represented by those data. Therefore, while ML shows great potential to complement traditional approaches in geotechnical engineering, it is crucial to address these limitations before the implementation of any ML model.

Recently, numerous studies have employed ML algorithms to predict c_c from some parameters related to this property showing promising results (e.g. Desai et al. 2009; Kalantary and Kordnaeij 2012; Kumar and Rani 2011; Nesamatha and Arumairaj 2015; Samui et al. 2012). Alam et al. (2014) compiled a database of 125 clay samples, which included w, LL, e_0, and PI as input variables and created an Artificial Neural Network (ANN) model to predict c_c. Kumar and Rani (2011) also used an ANN to predict c_c, considering 41 samples with the following input variables: fine contents, LL, PI, maximum dry density, and optimal moisture content. Park and Lee (2011) developed an artificial neural network (ANN) model by utilizing 947 consolidation tests

84 performed on soil samples gathered from 67 construction sites in the Republic of Korea. They
85 considered as input variables, w, LL, PI, e_0, specific gravity of soil particles (G_s), and weight
86 percentage of sand, silt and clay. Benbouras et al. (2019) developed an ANN model with 373
87 oedometer test samples to correlate c_c with wet density, w, e_0, fine content, LL, PI, and soil type.
88 The dataset utilized in this research comprised samples gathered from various projects executed
89 in the city of Algiers (Algeria). Zhang et al. (2021) used a Random Forest algorithm that utilized
90 a database with 311 samples, encompassing three input variables (LL, PI, e_0). Asteris et al. (2022)
91 introduced extreme learning machine models that applied Manta ray foraging optimization to
92 predict c_c from void ratio at liquid limit, LL and PI. It should be noted that the void ratio at LL is
93 a parameter that is not usually available in the design phases of geotechnical projects, whereas
94 e_0, is more frequently encountered. Long et al. (2023) established a relationship between c_c and
95 w, LL, PI, e_0, and G_s using Tree-Based Techniques from 391 samples from Northern Iran.
96 However, all these approaches have, to a greater or lesser extent, a local character, or a relatively
97 small number of samples. All the studies presented either rely on a limited dataset, or the collected
98 samples have a local nature (i.e., they come from the same area or country), or they do not have
99 a direct application because they are based on ML algorithms. Many engineers lack knowledge
100 of the programming language or specific software required to utilize the proposed algorithms.
101 Therefore, the main objective of this work is threefold: 1) to create an algorithm based on a dataset
102 with a large number of samples, 2) to ensure these samples are collected worldwide to eliminate
103 any local bias, and 3) to design the proposed algorithm in a way that any geotechnical engineer
104 can easily use it. The present paper took advantage of ensemble learners, to develop and validate
105 an accurate prediction model for forecasting the compression index of clay soils for a large
106 number of data (1008) collected worldwide. Subsequently, this model was deployed to a user-
107 friendly web application. This paper is structured as follows. A brief summary of the experimental
108 database is given in Section 2. The machine learning process performed is presented in Section
109 3. Later, in Section 4, the main results obtained are studied and interpreted by evaluating the
110 super-learner machine learning model. Section 5 describes the online tool deployed. An analysis

of the dataset using Symbolic Regression is included in Section 6. The final section (Section 7) ends the paper and presents the main conclusions obtained.

2. Database description and analysis.

The collection of the dataset is the first step in the building of a machine learning model. In the present work, the experimental database of c_c (1008 samples) comprised 913 samples from research papers (Alhaji et al. 2017; Benbouras et al. 2019; Kalantary and Kordnaeij 2012; LCPC 1977; McCabe et al. 2014; Mitachi and Ono 1985; Widodo and Ibrahim 2012), and 95 samples from authors' own data. This database has samples of different countries as Nigeria, Ireland, Spain, Iran, Indonesia, France, Algeria, Bangladesh, among others. In the same way, the database has soils with low plasticity to very high plasticity. In fact, LL ranges from 17.1% to 199.0% and PI from 2.0% to 82.0%. On the other hand, the database included soils with very high compressibility and soils with low compressibility according to e_0 values (ranging from 0.279 to 7.114). This fact is also corroborated, by examining the values of c_c varying these between 0.013 and 2.2. Finally, the w values, also varies widely ranging from 8% to 244.1%. Based on all the above, the database includes a significant amount of data and a wide range of them, making it suitable for a reliable study. The final dataset can be found in Table S1 (accessible online), which includes details such as references, soil type, mineralogy, and origin when available. Furthermore, the dataset includes both the actual c_c values and those obtained by the super-learner ML model. The main statistics of the compiled database are shown in Table 1.

	LL (%)	**PI (%)**	e_0	w **(%)**	c_c
Samples	1008	947	1008	1008	1008
Mean	46.49	21.78	0.817	29.73	0.236
Standard deviation	15.90	9.83	0.467	17.06	0.210
Minimum	17.10	2.00	0.279	8.00	0.013
25th percentile	36.00	15.07	0.598	21.00	0.143
Median	44.85	22.03	0.712	25.40	0.180
75th percentile	54.14	28.05	0.852	31.20	0.246
Maximum	199.00	82.00	7.114	244.10	2.200

Table 1. Descriptive statistics of the data analysed.

Figure 1 shows box plots for LL, PI, e_0, w and c_c of the database. Box plots are used to visualize data dispersion, which was split into quartiles. The method is used to detect outliers (if any), data symmetry, dispersion, and skewness (Reagan and Kiemele 2008). The box in a box plot shows the interquartile range (IQR), with the bottom and top of the box representing the 25th and 75th percentiles, respectively. The whiskers extend to the final data value inside the inner fence, which is 1.5 times the IQR from the box's edge. The height of the box represents the interquartile range. Outliers are defined as data points extending to 3×IQR (Reagan and Kiemele 2008). Some points in Figure 1 are identified as outliers, mainly corresponding to high values of the variables. The analysis of these box plots reveals the wide range of variation and high dispersion of the variables studied clearly related to the worldwide character of the compiled database. For a better interpretation of Figure 1, the data included in Table 1 can be consulted.

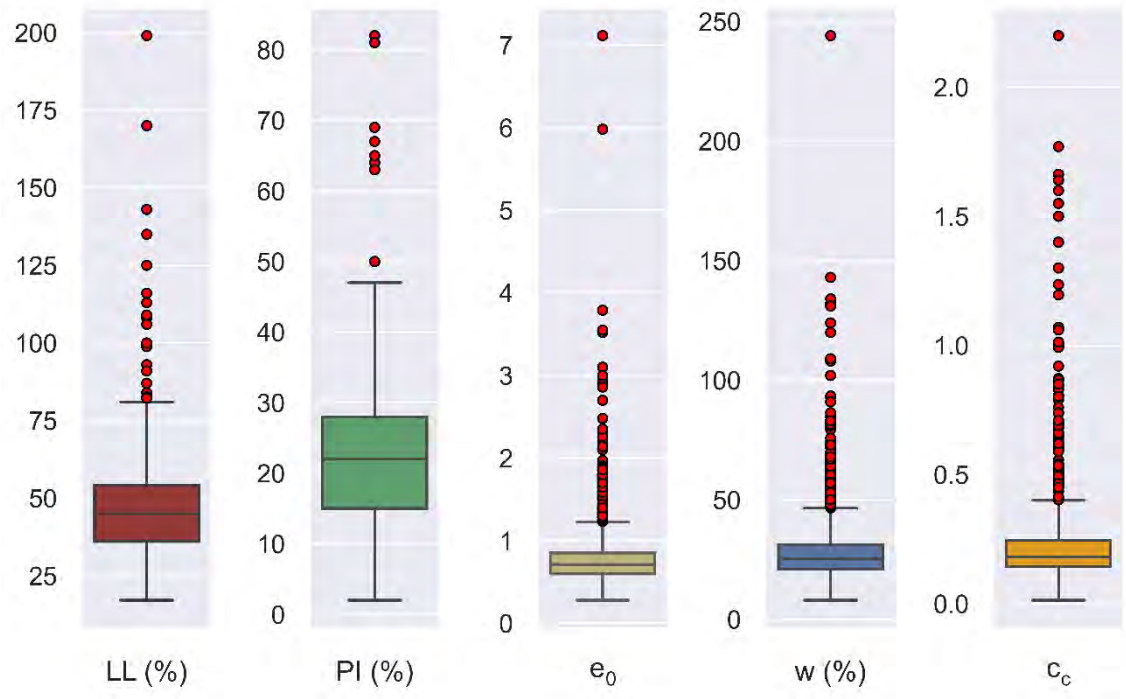


Figure 1. Box plots of the variables considered.

The scatter plots of the considered parameters are depicted in Figure 2, to show a descriptive overview of the data distribution. These plots indicate a positive relationship between c_c and the rest of the variables. This fact implies, to a greater or lesser extent, that an increase in the

considered input variables tends to proportionally increase c_c. This relationship is clearer in variables as e_0 and w and more diffuse in LL and PI. Finally, a strong relationship between e_0 and w must also be noted.

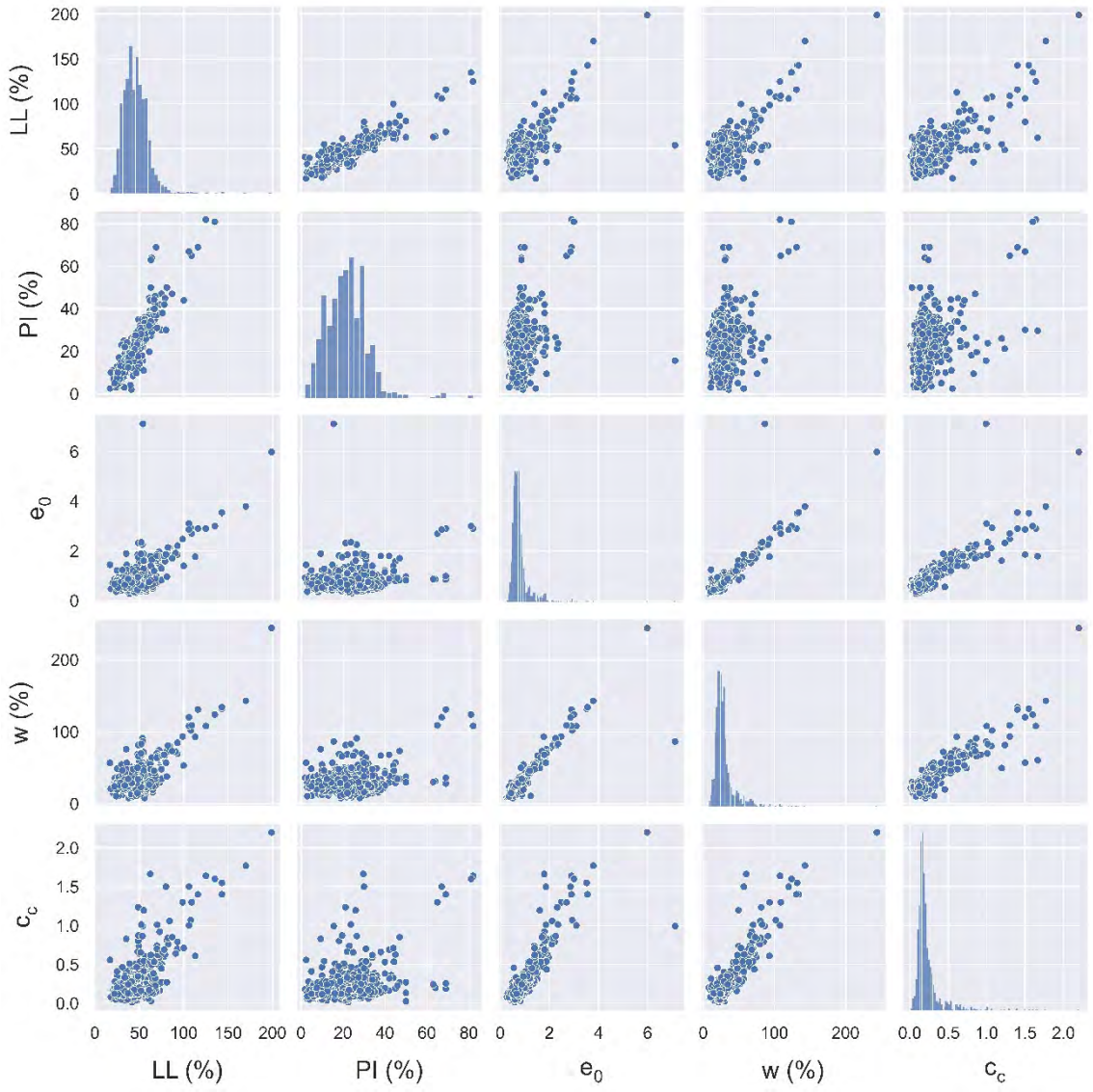


Figure 2. Scatter plots and distribution histograms of the variables.

These linear trends can be quantified numerically using the Pearson correlation coefficient (r) and a correlation matrix. A correlation matrix is a table showing the correlation values, which measure the degree of linear relationship between each pair of variables. Correlation values can be between -1 and +1. If the two variables tend to increase or decrease at the same time, the correlation value

is positive. In Table 2 is shown the correlation matrix providing an overview of the Pearson correlation coefficient. From the analysis of the table, it can be verified that the same relationships established previously in a visual manner are the ones that obtain the highest correlation values. Indeed, a strong and positive correlation is shown between c_c and e_0 (r=0.87) and w (r=0.89), and a somewhat less strong and also positive correlation between c_c and LL (r=0.66). Between c_c and PI, the correlation is low and positive (r=0.34). On the other hand, there is a strong and positive correlation between e_0 and w (r=0.92).

	LL	**PI**	$\boldsymbol{e_0}$	$\boldsymbol{w}$	$\boldsymbol{c_c}$
LL	1.00				
PI	0.92	1.00			
$\boldsymbol{e_0}$	0.61	0.23	1.00		
$\boldsymbol{w}$	0.64	0.28	0.92	1.00	
$\boldsymbol{c_c}$	**0.66**	**0.34**	**0.87**	**0.89**	**1.00**

Table 2. Correlation matrix of the variables considered. Note that the values indicated correspond to the Pearson correlation coefficient (r).

The equations obtained by linear regression for the cases with the highest value of Pearson correlation coefficient are (Equations 1 to 3):

$$c_c = 0.393 e_0 - 0.0842 \quad (1)$$

$$c_c = 0.011 w(\%) - 0.0905 \quad (2)$$

$$e_0 = 0.0252 w(\%) + 0.0687 \quad (3)$$

The coefficient of determination (R^2) of each of the three equations above presented is 0.76, 0.80, 0.85, respectively. Equation 3 is very interesting, as the e_0 value can be estimated by w, considering that e_0 relates the soil structure with its geologic history (Onyejekwe et al. 2015).

3. Machine learning procedure.

3.1. Model selection process.

According to the analysis of the dataset carried out in the previous section, PI has 61 data less than the rest of the variables. To match the number with the rest of the variables, a data imputation technique was used. This is a technique widely used in ML algorithms for dealing with missing values and it has been used in geotechnical issues satisfactorily (e.g. Aydın et al. 2023; Díaz et al. 2023). For the imputation of values, a multivariate feature imputation algorithm has been chosen (Little and Rubin 2019; Van Buuren and Oudshoorn 2000). This technique uses the information of all of the available features in order to estimate the missing value of one variable by considering samples which have a similar situation in terms of all of the features in the dataset. After imputing the missing data, an outlier detection analysis was conducted using the one-class SVM algorithm developed by Schölkopf et al. (1999) and successfully employed in similar works (e.g. Díaz et al. 2023). This algorithm is a method designed to identify outliers and anomalies within a dataset, utilizing the principles of traditional Support Vector Machines (SVM). The core idea behind SVM is to identify a hyperplane that maximizes the separation between different classes within the dataset. Once this hyperplane is established, new data points can be classified based on their position relative to the hyperplane. In the case of the one-class SVM, there is only one class, and it defines the hyperplane for normal data points while classifying data points located outside the hyperplane as outliers. Upon applying this algorithm, 9 outliers were identified. These 9 data points were thoroughly examined, but no anomalous values in their properties were observed. Therefore, due to their limited number, it was decided not to remove them from the dataset. With the final dataset including 1008 records of each variable, a machine learning algorithm selection process was carried out using the k-fold cross-validation technique (with k=10). The results of this selection process are shown in Table 3, ordering the selected algorithms, from best to worst

performance considering four statistical indicators, mean absolute percentage error (MAPE), coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Three algorithms, all of them based on decision trees, outperform over the rest: Extra Trees Regressor (Geurts et al. 2006), Random Forest Regressor (Ho 1995), and Gradient Boosting Regressor (Friedman 2001).

Model	RMSE	MAPE	MAE	R^2
Extra Trees Regressor	0.0767	0.2521	0.0466	0.87
Random Forest Regressor	0.0776	0.2534	0.0473	0.86
Gradient Boosting Regressor	0.0781	0.2529	0.0478	0.86
Huber Regressor	0.1050	0.2766	0.0559	0.75
Light Gradient Boosting Machine	0.1059	0.2831	0.0576	0.74
K Neighbors Regressor	0.1000	0.2919	0.0578	0.75
AdaBoost Regressor	0.0876	0.3427	0.0581	0.81
Ridge Regression	0.1084	0.3106	0.0610	0.74
Linear Regression	0.1117	0.3106	0.0613	0.72
Least Angle Regression	0.1117	0.3106	0.0613	0.72
Bayesian Ridge	0.1113	0.3124	0.0616	0.72
Decision Tree Regressor	0.0966	0.3296	0.0622	0.79
Elastic Net	0.1040	0.3307	0.0625	0.74
Orthogonal Matching Pursuit	0.1034	0.3471	0.0654	0.74
Lasso Regression	0.1170	0.3479	0.0668	0.68
Lasso Least Angle Regression	0.1170	0.3479	0.0668	0.68
Passive Aggressive Regressor	0.1232	0.3562	0.0728	0.63

Table 3. Performance of base machine learning models obtained by k-fold cross validation.

This process was carried out with the variables without normalization. But it should be noted that the same algorithm selection process was also performed normalizing the variables using the min-max method, which scales each variable individually between zero and one. The results of this process normalizing the variables, were exactly the same in the three models with the best performance.

3.2. Model development.

To ensure a proper generalization of the algorithms, it is good practice to assess their performance on unknown data. For this purpose, the dataset was divided into two groups (training and test) with an 80/20 partition. Subsequently, the three algorithms with the best performance were

subjected to a tuning process of their hyperparameters to maximize their performance. For this purpose, the particle swarm optimisation (Kennedy and Eberhart 1995) was used. This is a computational technique which optimises problems by iteratively improving candidate solutions (aka particles). The results of this optimization are shown in Table 4, in terms of R^2, RMSE, and MAE and with the result in both in the training and in the test set.

Model	Set	R^2	MAE	RMSE
Extra Trees Regressor	Training	0.92	0.034	0.059
	Test	0.92	0.040	0.055
Random Forest Regressor	Training	0.92	0.039	0.061
	Test	0.92	0.041	0.055
Gradient Boosting Regressor	Training	0.93	0.038	0.056
	Test	0.91	0.042	0.058

Table 4. Summary of performance metrics of the trained machine learning algorithms.

With the tuned algorithms, they were ensembled to improve their predictive capacity. Among the existing approaches, the super-learner algorithm (Laan et al. 2007) was chosen. The super-learner algorithm is a type of ensemble method that applies stacked generalization to k-fold cross-validation. It combines multiple prediction models (base learners) by assigning different weights to these models to find their optimal combination and produce a single best prediction function. Thus, the predictions of the base learners are used to train a regression model (meta learner) that assigns relative weights to the predictions of each base-model. In this case, and after a process of trial and error, the best results were obtained using the tuned Extra Trees Regressor and Gradient Boosting Regressor algorithms as base learners, and the tuned Random Forest Regressor algorithm as meta learner. It must be indicated that the meta learner was trained on the base models' predictions as well as the original training data. The performance metrics resulting in the super-learner machine learning model, are shown in Table 5.

Model	Set	R^2	MAE	RMSE
Super-Learner	Training	0.93	0.034	0.057
	Test	0.93	0.039	0.053

Table 5. Summary of performance metrics of the super-learner machine learning model.

As can be seen the super-learner machine learning model improved slightly the predictions of the best of the three selected algorithms (Table 4).

4. Results.

4.1. Analysis of the super-learner ML model for predicting c_c.

The prediction accuracy of the super-learner machine learning model is evaluated using training and test datasets. Figure 3 shows the scatter plots for the actual values of c_c (x-axis) versus the predicted values of c_c (y-axis). This figure shows that the vast majority of data are located close to the no error line, indicating a good agreement between the predicted and measured values. Additionally, in Figure 3 are included the values of R^2 on both the training and test datasets. These values (0.93 in both sets) are indicative of a high predictive capacity, and since they are the same, it is assured a correct behaviour of the model in unseen data, discarding overfitting issues. The predictive performance of the selected model was also evaluated based on the anteriorly defined performance metrics (MAE and RMSE). In the training dataset, MAE and RMSE values of 0.034 and 0.057 were obtained, respectively. In the test set, these values were similar (MAE of 0.039 and RMSE of 0.053). These metrics are summarized in Table 5 (previous section) for both the train and test datasets. Additionally, the a20-index (Apostolopoulou et al. 2020; Asteris et al. 2021a; Asteris et al. 2021b) was calculated. The a20-index offers the advantage of having a clear engineering interpretation, indicating the percentage of samples that meet the predicted values within a ±20% deviation from the experimental ones. The values obtained in the training and test sets were 73.45% and 71.27%, respectively, showing a similar and reasonable performance for the prediction of c_c.

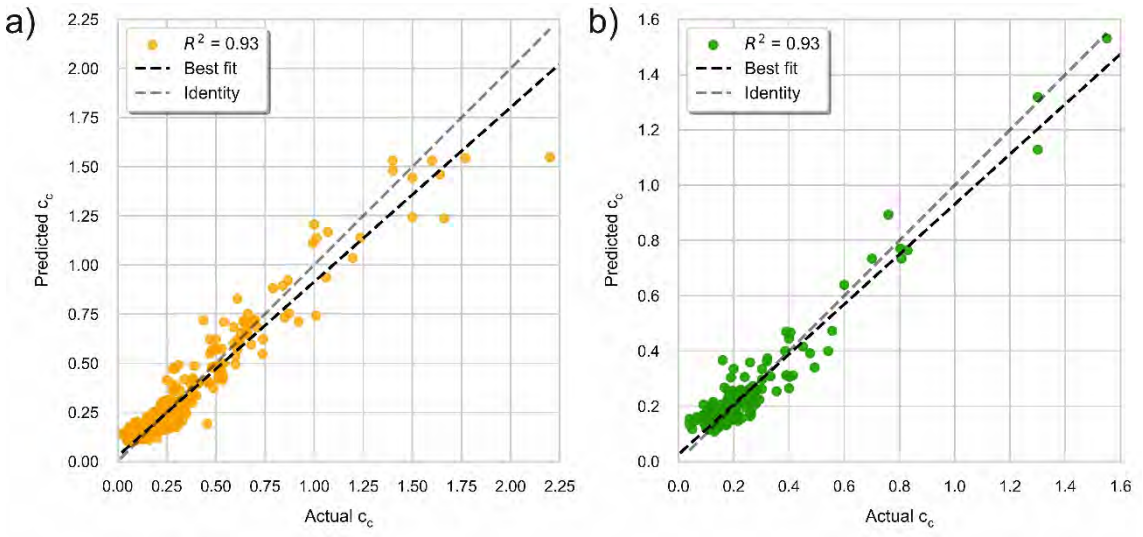


Figure 3. Scatter plot showing graphical performance of the super-learner machine learning model in (a) training set, and (b) test set. The identity line (i.e., no error line) and the best fit line were included in both figures.

In order to evaluate the results of the adopted model, a residuals study was performed. The residuals were defined as the difference between the predicted and actual value obtained by the super-learner model. In Figure 4 are shown the residuals for both the training and test datasets with two different visualizations, a scatter plot, and a histogram distribution. An inspection of this figure shows that for the two datasets considered, the residuals are concentrated around zero and have Gaussian distributions (i.e., residuals are normally distributed) and, thus, there are not many outliers. These facts are indicative of a robust algorithm with a great generalization ability and do not reveal significant issues in the predictions.

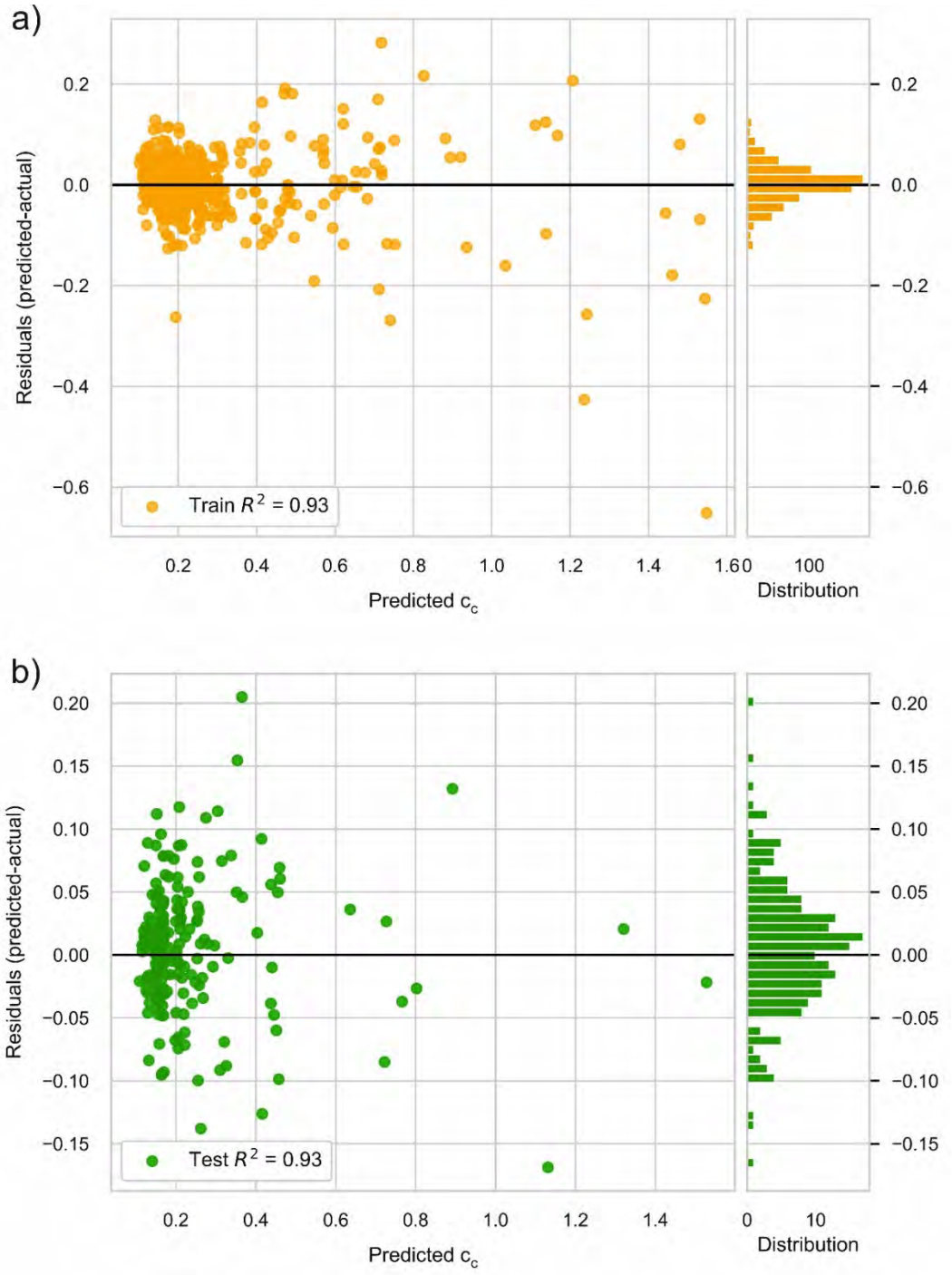


Figure 4. Residuals scatter plot and histogram distribution in (a) training set and (b) test set of the super-learner machine learning model.

Figure 5 compares the experimental and predicted c_c values on the test set (unseen data). The prediction performance of the testing set, according to what was previously discussed, was very good. Figure 5 shows that the super-learner machine learning model, can accurately capture the

evolution of the actual values, the prediction results are consistent, and the difference between the minimum and maximum predicted values is relatively small. No large deviations are observed in a general manner, affecting only isolated samples.

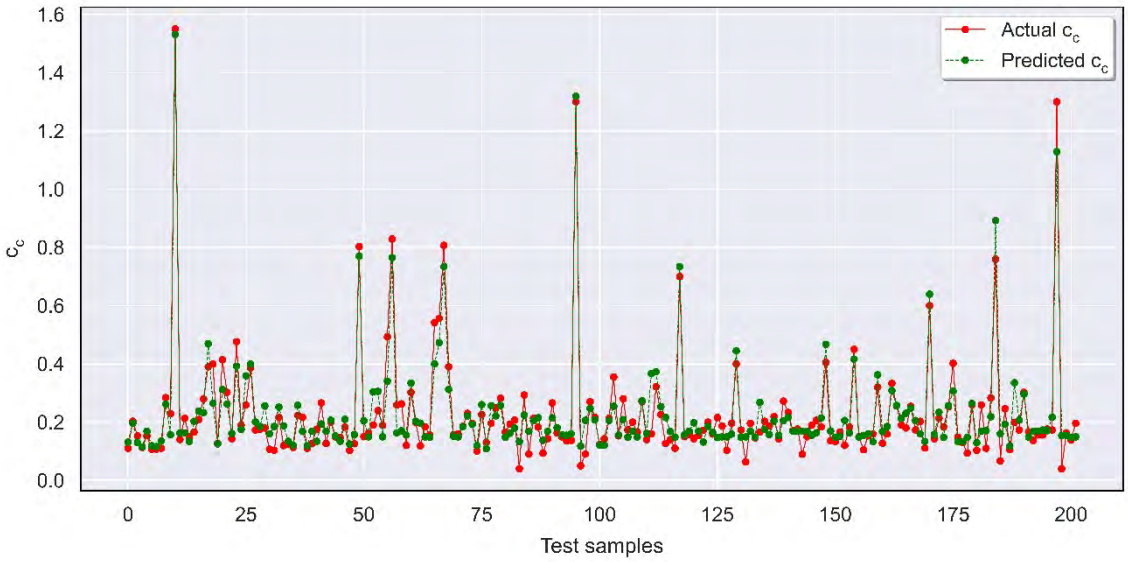


Figure 5. Performance plot of super-learner machine learning model prediction results vs actual values of c_c in the test set.

4.2. Sensitivity analysis.

Subsequently, a sensitivity analysis of the parameters involved in the super-learner ML model was conducted. This analysis employed Sobol's method (Sobol 1990), a technique for assessing the significance of input parameters in computational models. Sobol's method evaluates their impact on the output by quantifying their contribution to output variance. Within this method, the First-order sensitivity indices (S1) measure the influence of each variable in isolation while holding all other variables constant. Total-order sensitivity indices (ST) encompass not only the individual effects of each variable but also their interactions with other variables on the output. The results of this analysis are collected in Table 6, showing that e_0 was the most influential variable, with a ST of 0.839, implying that it had the greatest effect on the variability of the model's output. Following e_0, w emerged as the second most important factor in this analysis (ST = 0.122). Next, LL possessed a moderate ST, with a value of 0.066. Finally, PI exhibited the

lowest ST, with a value of 0.018. The values of ST and S1 for each variable were similar, suggesting that the majority of the influence of the variables on the model's output was attributed to their individual effects rather than complex interactions among them.

Variable	ST	S1
LL	0.066	0.064
PI	0.018	0.016
$\boldsymbol{e_0}$	0.839	0.802
$\boldsymbol{w}$	0.122	0.080

Table 6. Sensitivity analysis of variables using Sobol's method comparing the Total (ST) and First-Order (S1) Sensitivity Indices for input variables.

After establishing the significance of the variables, an analysis was conducted to understand the impact of variations in these inputs on the overall results. Initially, partial dependence plots for single input features were analysed to visualize the relationship between individual input variables and the output variable in a regression model. These plots are instrumental in understanding how a single input variable influences the output while holding other variables constant. The results of this analysis are displayed in Figure 6 in terms of partial dependence (average predicted effect). For the variables LL and PI, an increasing trend was observed, with some fluctuations, indicating a positive and near linear relationship with c_c. In the case of variable e_0, a very strong growth trend was noted until e_0=2.5, after which the impact stabilizes. This suggests that within this range of values (up to e_0=2.5), small variations in e_0 will cause a significant impact on the value of c_c, clearly indicating that the relationship is not linear. As for w, a similar pattern to e_0 was observed, but with a lesser impact on the value of c_c, with the effect becoming stable around values of 110%.

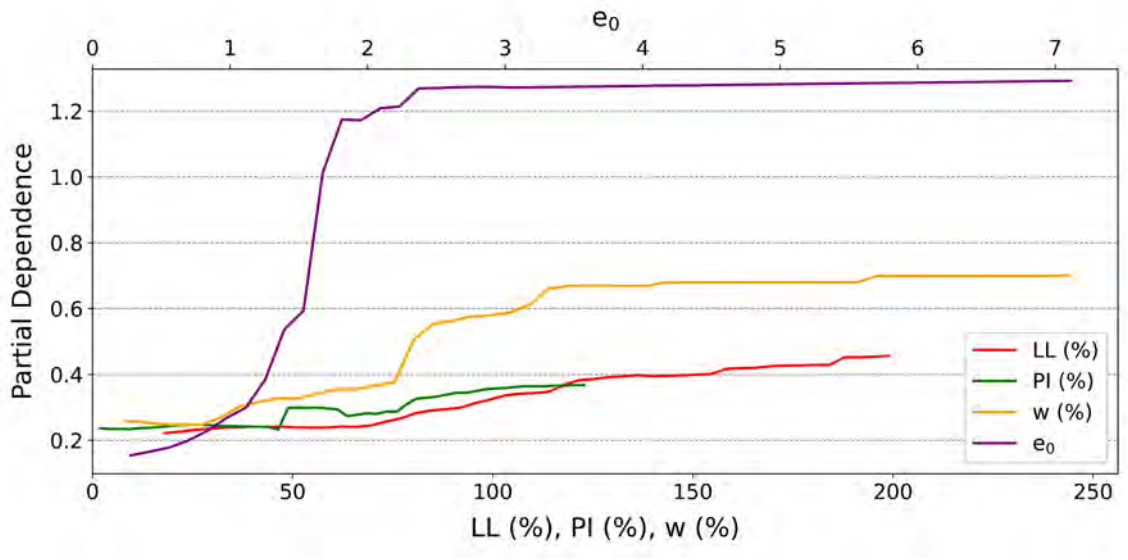


Figure 6. Partial dependence plots for single input features in terms of the average predicted effect. Note that the figure has a dual x-axis, with the upper one for e_0 and the lower one for the variables LL, PI and w.

Finally, partial dependence plots for pairs of input features were computed. These plots are designed to illustrate the joint influence of input feature pairs on the model's predictions, offering insights into their collective impact on the output variable. They assist in elucidating complex interactions between input features and the model's response. These graphs depicting the pairs of variables LL and PI, as well as w and e_0, are presented in Figure 7. The graphs showing the remaining variable pairs are included in Figure S1. From this analysis, it was established that most of the impact on the model's output can be attributed to the individual effects of the variables, rather than complex interactions among them. In the LL and PI pair (Figure 7a), as well as in the w and e_0 pair (Figure 7b), some interaction was observed, although not excessively significant. These conclusions were consistent with the Sobol analysis.

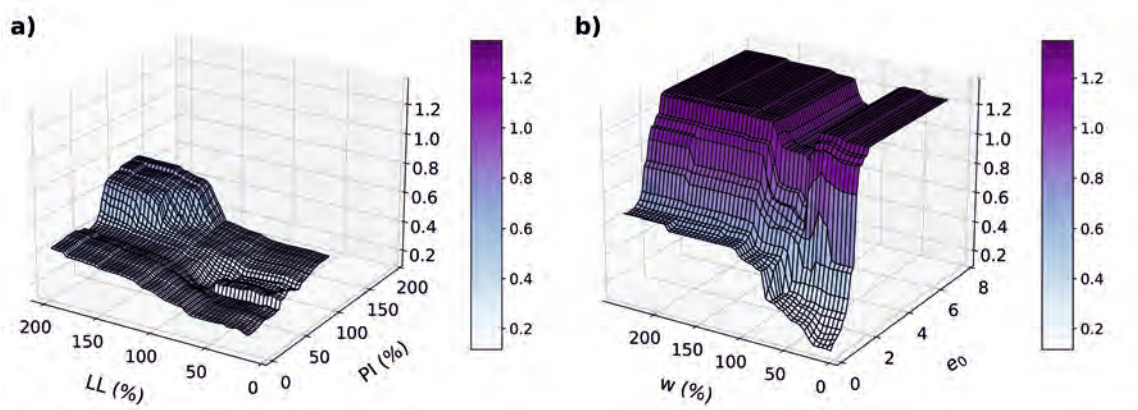


Figure 7. Partial dependence plots for pairs of input features. a) LL - PI and b) w - e_0.

In conclusion, the analyses conducted highlighted that the most significant variable in the model was e_0, which also caused the most substantial changes in the predicted value of c_c, both individually and in terms of interactions between pairs of variables.

4.3. Overfitting analysis

One of the common challenges encountered in ML models is overfitting. This occurs when a model performs exceptionally well in replicating the data used for its development and training. However, when applied to input parameter values outside those used during development and training, the model may generate highly unusual predictions. To analyse the overfitting of a ML model, there are several techniques, highlighting the works of Asteris et al. (2019) or Armaghani and Asteris (2021), where the final model is checked with datasets in which the inputs gradually increase in value as they are tests made for a subsequent analysis. In this study, it is not like this since the inputs do not increase gradually as they were tests conducted on soils where the values of the input variables do not follow a gradual pattern. Thus, alternative techniques were chosen to enable the identification of overfitting in a model. To this end, the difference in model performance between the training set and the test set was initially assessed. In this case, a very small difference was observed between the training and test errors (Table 5), indicating that there was no evidence of overfitting. Alternatively, analysing the residuals can provide insights into potential overfitting in the model. Specifically, the residuals of the train and test sets were

analysed based on the following criteria: a) size: if the test residuals are significantly larger than the train residuals, it is a sign of overfitting; b) variance: if similar variances exist in train and test, it is indicative that the model generalizes well, not evidencing overfitting; and c) correlation: if the residuals are correlated, it is a sign that the model is learning patterns that are not relevant for predicting the dependent variable, which is a sign of overfitting. In this case, the size of the residuals and their variance were given by the mean and standard deviation, respectively. The correlation of the residuals was carried out through the Durbin-Watson test, which is a method to detect the presence of autocorrelation among the residuals in regression analysis. The value of the Durbin-Watson statistic varies between 0 and 4. A value close to 2 suggests that there is no autocorrelation; values less than 2 indicate positive autocorrelation; and values greater than 2 indicate negative autocorrelation. The summary of this analysis is shown in Table 7.

Set	Mean	Standard deviation	Durbin–Watson statistic
Training	0.0003	0.05730	1.95
Test	−0.0038	0.05301	1.87

Table 7. Summary of residual metrics in the training and test sets.

From the analysis of the results, it was evident that the mean of the residuals is very close to zero for both sets, which is good as it suggested that there is no systematic bias in the predictions. The standard deviations were also comparable, indicating that the dispersion of errors was similar in both sets. The Durbin-Watson values were close to 2, implying that the residuals exhibited independence. According to the results obtained in the residual analysis, there is no evidence of overfitting from this perspective.

Although none of the conducted analyses indicated signs of overfitting in the model, an additional analysis was undertaken by perturbing the input values of the test set. In this analysis, a random perturbation of ±1% was applied to the test set data (including all input variables) to simulate natural variability and assess the robustness of the model and its propensity for overfitting. Subsequently, the super-learner ML model was tasked with making predictions on both the original test set and the perturbed test set to evaluate how the model reacted to disturbances in the

input data. A model that is excessively sensitive to small perturbations in the data might be overfitting, meaning it has learned to adjust to the particularities and noise of the training set, rather than capturing general trends. The results of this analysis are presented in Figure 8, where the scatter plot revealed an almost perfect correlation (R^2 of 0.999) between the original and perturbed predictions. The histogram showed that the average changes in predictions due to perturbations were minimal, with a mean of 0.0036, suggesting significant stability of the model against minor variations in the data. As a conclusion of this analysis, the model was highly robust and showed low sensitivity to disturbances in the input data. Therefore, the three conducted analyses concluded that the super-learner ML model exhibited strong generalization abilities, capturing underlying trends rather than fitting to specific noises or peculiarities of the training data.

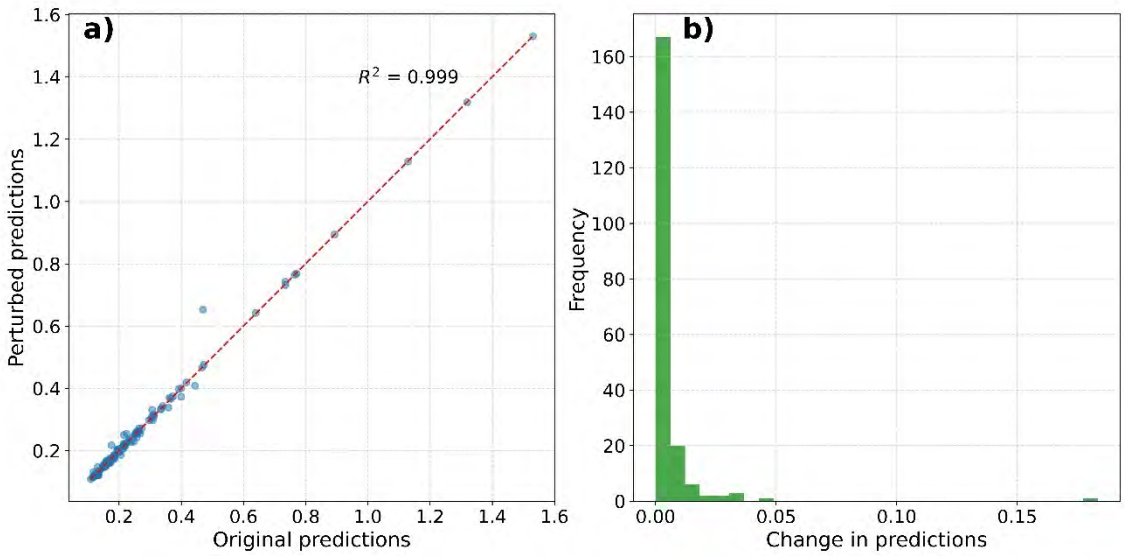


Figure 8. Analysis of model robustness to input perturbation. a) Scatter plot comparing original and perturbed model predictions, and b) Histogram of absolute changes in model predictions due to perturbation.

4.4. Comparison of the super-learner ML model with geotechnical correlations.

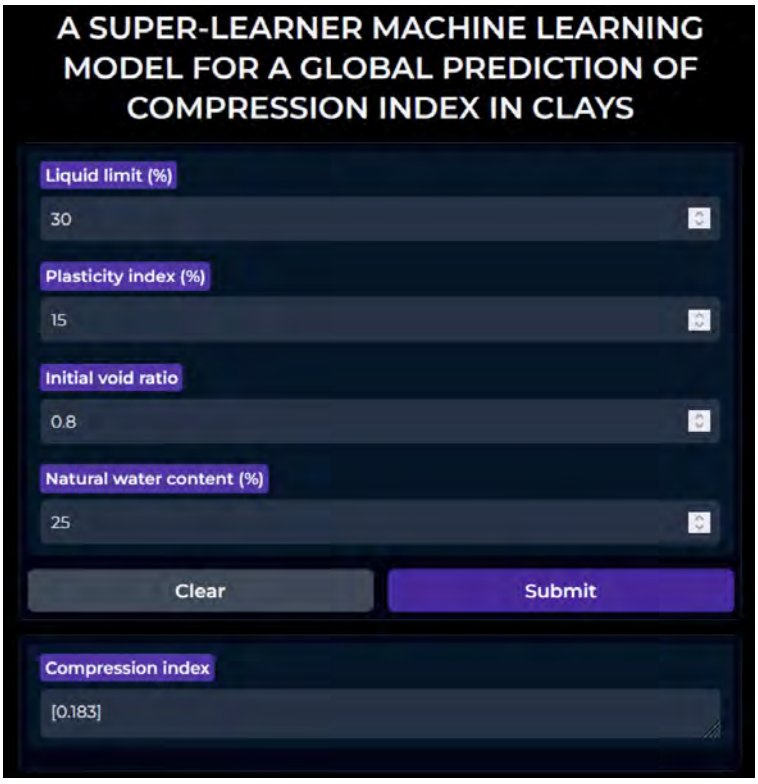
In order to assess the performance of the proposed model, a comparative analysis was conducted using the database compiled for this study. This analysis involved nine empirical correlations for

estimating c_c with a global scope. The results are depicted in Figure S2 as a scatter plot, showcasing the comparison between each correlation's predictions and the actual values from the comprehensive dataset. Additionally, in this figure, the formulas of the tested correlations are included. The super-learner ML model proposed in this work clearly achieved the best performance overall, with the highest R^2 score, lowest RMSE, and lowest MAE (Figure S2a). In contrast, the Sridharan and Nagaraj (2000) correlation (Figure S2b) performed very poorly, with a negative R^2 score (-0.79), high RMSE (0.281), and high MAE (0.257), clearly showing it did not work well for this dataset. Other correlations such as Azzouz et al. (1976) (Figure S2e), Rendon Rendon-Herrero (1980) (Figure S2g), and Koppula (1981) (Figure S2f) achieved moderate results, with R^2 scores around 0.7, RMSE values around 0.10, and MAE values around 0.06. The Wroth and Wood (1978) (Figure S2c), Bowles (1979) (Figure S2h), and Al-Khafaji and Andersland (1992) (Figure S2i) correlations demonstrated low performance, characterized by low R^2 scores around 0.20, RMSE values around 0.19, and high MAE values. Additionally, the Koppula (1981) (Figure S2j) correlation also had poor outcomes, with negative R^2 scores, high RMSE, and high MAE. It is noteworthy that some correlations, such as Sridharan and Nagaraj (2000) (Figure S2b and d), Koppula (1981) (Figure S2f and j), and in less extent, Wroth and Wood (1978) (Figure S2c), consistently predicted values significantly above the actual ones. Conversely, correlations like Al-Khafaji and Andersland (1992) (Figure S2i) and Bowles (1979) (Figure S2h) tended to predict consistently below the actual values. Furthermore, the four expressions obtained through Symbolic Regression in this study (Equations 4 to 7) also exhibited superior performance metrics compared to the analysed correlations. In summary, the model proposed in this paper was unequivocally the most accurate and effective for this dataset.

5. Online tool developed

Machine learning models have seen their use increase exponentially in recent years due to their high performance on the predictions and for their ability of discovering robust patterns in complex datasets. However, their application is usually difficult because it is necessary to know the trained algorithm fully, with all its hyperparameters and to know the tools with which it has been

444 developed (e.g., programming language or specific software). This means that currently, many
445 machine learning researches do not have a great practical application. In this study, an online
446 application has been developed to facilitate the non-expert user, the usage of the machine learning
447 model presented here, becoming one of the few studies that offers a practical implementation of
448 the trained machine learning algorithm. Figure 9 shows the graphical user interface of the
449 developed prediction tool, which is located at
450 https://huggingface.co/spaces/EstebanDC/Compression_Index. In this tool, the value of c_c of a
451 clay soil can be easily obtained by the user, defining the input values (i.e., LL(%), (PI %), w(%)
452 and e_0). Once these values are introduced, the compression index is calculated using the optimized
453 super-learner ML model. This online tool can be used on any device with an internet connection
454 (computers, tablets, or smartphones).
455



456

457 *Figure 9. Application developed with the trained super-learner machine learning model.*

458

459

460

6. Symbolic regression

In addition to the analysis carried out using the machine learning models presented previously, the dataset was analysed using Symbolic Regression. The aim of this analysis is to propose mathematical expressions, which relate c_c with some of the input variables. These expressions can be applied alternatively to the super-learner machine learning model when data for all the input variables are not available. Symbolic Regression-based approaches provide alternative machine learning methods that are recently gaining popularity. These approaches, as other ML algorithms, learn patterns from observed data and they have a great advantage over other ML algorithms due to their interpretability and explanation capabilities. Symbolic Regression attempts to explain a target variable by multiple input variables using a mathematical expression involving of a predefined set of basic computation functions. In this study, a symbolic regressor based on Feynman's path integral formulation from quantum field theory (Broløs et al. 2021) was applied. With this approach, four expressions (Equations 4 to 7) were obtained, and their performance metrics are presented in Table 8.

$$c_c = 1.95 - 2.04\tanh(1.82(0.204e_0 - 1)^2) \quad (4)$$

$$c_c = 1.72e^{-4.02(1-0.209e_0)^4 - 0.599(0.00363LL-1)^2} + 0.0513 \quad (5)$$

$$c_c = 3.32e^{-3.82(1-0.219e_0)^2 - 2.0e^{-9.78(1-0.244e_0)^2 - 0.237(-0.0135PI-1)^2}} - 0.0409 \quad (6)$$

$$c_c = 2.67e^{-1.99(1-0.209e_0)^2 - 0.406(0.00286w-1)^2} - 0.26 \quad (7)$$

With w, LL, and PI expressed in %.

Equation	Set	R^2	MAE	RMSE
4	Training	0.85	0.049	0.083
	Test	0.87	0.042	0.065
5	Training	0.86	0.047	0.072
	Test	0.86	0.049	0.088
6	Training	0.86	0.047	0.078
	Test	0.91	0.043	0.063
7	Training	0.85	0.048	0.082
	Test	0.92	0.047	0.068

Table 8. Summary of performance metrics of the proposed equations.

Symbolic Regression approach gets good prediction performance, albeit being less accurate than the trained super-learner machine learning model, although presenting moderately long training times. Alternatively to the Equations 4 to 7, those obtained by linear regression and included in section 2 (Equations 1 to 3) can be used, although they offer worse metrics than those obtained by Symbolic Regression. These expressions must be fed with values of the input variables within the ranges with which the algorithm has been trained, since the behaviour outside these ranges is unknown.

7. Conclusions.

This study proposes a novel super-learner ML algorithm that provides a reliable and accurate model for predicting c_c, a key parameter in engineering applications. This algorithm is reasonably capable of predicting the value of c_c based on variables previously related to this parameter (i.e., LL, PI, e_0 and w). To this aim, a database was built including more than 1000 samples collected worldwide, in order to reduce the local character of the proposals presented. The high and satisfactory error metrics, including R^2, MAE, RMSE, and a20-index, demonstrate the model's robust performance, particularly considering the global nature of the dataset and the dispersion in the input variables. To facilitate the use of the model, a user-friendly prediction tool has been deployed online (https://huggingface.co/spaces/EstebanDC/Compression_Index). Additionally, four symbolic regression expressions have been proposed for cases where all input parameters are not available. A comparative analysis was carried out using the compiled database against

global empirical correlations, demonstrating that super-learner ML model was the most accurate and effective for this dataset. The global scope of this study and the development of an online tool enhance the proposed algorithm's applicability and value beyond other methods in the field, thereby making it more useful for geotechnical engineers. Future work will focus on expanding the dataset and updating the model to increase its generalization ability and application range. Finally, it is important to exercise caution when applying the model or expressions beyond the specified input variable ranges, as this can lead to predictions of questionable reliability, a common issue with predictive algorithms.

Availability statement

The data that support the findings are available from the corresponding author upon reasonable request.

References

Al-Khafaji, A. & Andersland, O. 1992. Equations for compression index approximation. Journal of Geotechnical Engineering, **118**, 148-153, doi: https://doi.org/10.1061/(ASCE)0733-9410(1992)118:1(148).

Alam, S., Khuntia, S. & Patra, C. 2014. Prediction of compression index of clay using artificial neural network. *International conference on industrial engineering science and applications-NIT, Durgapur*.

Alhaji, M.M., Alhassan, M., Tsado, T.Y. & Mohammed, Y.A. 2017. Compression Index Prediction Models for Fine-grained Soil Deposits in Nigeria. Proceeding of the 2nd International Engineering Conference, Federal ….

Apostolopoulou, M., Asteris, P.G., Armaghani, D.J., Douvika, M.G., Lourenço, P.B., Cavaleri, L., Bakolas, A. & Moropoulou, A. 2020. Mapping and holistic design of natural hydraulic lime mortars. Cement and Concrete Research, **136**, 106167, doi: https://doi.org/10.1016/j.cemconres.2020.106167.

Armaghani, D.J. & Asteris, P.G. 2021. A comparative study of ANN and ANFIS models for the prediction of cement-based mortar materials compressive strength. Neural Computing and Applications, **33**, 4501-4532.

Asteris, P.G., Armaghani, D.J., Hatzigeorgiou, G.D., Karayannis, C.G. & Pilakoutas, K. 2019. Predicting the shear strength of reinforced concrete beams using Artificial Neural Networks. Computers and Concrete, An International Journal, **24**, 469-488.

Asteris, P.G., Koopialipoor, M., Armaghani, D.J., Kotsonis, E.A. & Lourenço, P.B. 2021a. Prediction of cement-based mortars compressive strength using machine learning techniques. Neural Computing and Applications, **33**, 13089-13121, doi: 10.1007/s00521-021-06004-8.

Asteris, P.G., Lourenço, P.B., Hajihassani, M., Adami, C.-E.N., Lemonis, M.E., Skentou, A.D., Marques, R., Nguyen, H., Rodrigues, H. & Varum, H. 2021b. Soft computing-based models for the prediction of masonry compressive strength. Engineering Structures, **248**, 113276, doi: https://doi.org/10.1016/j.engstruct.2021.113276.

Asteris, P.G., Mamou, A., Ferentinou, M., Tran, T.-T. & Zhou, J. 2022. Predicting clay compressibility using a novel Manta ray foraging optimization-based extreme learning machine model. Transportation Geotechnics, **37**, 100861, doi: https://doi.org/10.1016/j.trgeo.2022.100861.

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., Nigdeli, S.M. & Geem, Z.W. 2023. Use of Machine Learning Techniques in Soil Classification. Sustainability, **15**, 2374.

Azzouz, A.S., Krizek, R.J. & Corotis, R.B. 1976. Regression analysis of soil compressibility. Soils and Foundations, **16**, 19-29, doi: https://doi.org/10.3208/sandf1972.16.2_19.

Baghbani, A., Choudhury, T., Costa, S. & Reiner, J. 2022. Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. Earth-Science Reviews, **228**, 103991.

Bardhan, A., Alzo'ubi, A.K., Palanivelu, S., Hamidian, P., GuhaRay, A., Kumar, G., Tsoukalas, M.Z. & Asteris, P.G. 2023. A hybrid approach of ANN and improved PSO for estimating soaked CBR of subgrade soils of heavy-haul railway corridor. International Journal of Pavement Engineering, **24**, 2176494, doi: 10.1080/10298436.2023.2176494.

Bardhan, A., Gokceoglu, C., Burman, A., Samui, P. & Asteris, P.G. 2021. Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. Engineering Geology, **291**, 106239, doi: https://doi.org/10.1016/j.enggeo.2021.106239.
Benbouras, M.A., Kettab Mitiche, R., Zedira, H., Petrisor, A.-I., Mezouar, N. & Debiche, F. 2019. A new approach to predict the compression index using artificial intelligence methods. Marine Georesources & Geotechnology, **37**, 704-720.
Bowles, J.E. 1979. Physical and geotechnical properties of soils.
Broløs, K.R., Machado, M.V., Cave, C., Kasak, J., Stentoft-Hansen, V., Batanero, V.G., Jelen, T. & Wilstrup, C. 2021. An approach to symbolic regression using feyn. arXiv preprint arXiv:2104.05417.
Carter, M. & Bentley, S.P. 1991. Correlations of soil properties. Pentech press publishers.
Craig, R.F. 2004. Craig's soil mechanics. CRC press.
Dam Nguyen, D., Roussis, P.C., Thai Pham, B., Ferentinou, M., Mamou, A., Quang Vu, D., Thi Bui, Q.-A., Kien Trong, D. & Asteris, P.G. 2022. Bagging and Multilayer Perceptron Hybrid Intelligence Models Predicting the Swelling Potential of Soil. Transportation Geotechnics, **36**, 100797, doi: https://doi.org/10.1016/j.trgeo.2022.100797.
Das, B.M. 2021. Principles of geotechnical engineering. Cengage learning.
Desai, V.M., Desai, V. & Rao, D. 2009. Prediction of compression index using artificial neural networks. *Indian geotechnical conference (IGC-2009), Guntur, India*, 614-617.
Díaz, E., Salamanca-Medina, E.L. & Tomás, R. 2023. Assessment of compressive strength of jet grouting by machine learning. Journal of Rock Mechanics and Geotechnical Engineering, doi: https://doi.org/10.1016/j.jrmge.2023.03.008.
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.
Friedman, J.H.J.A.o.s. 2001. Greedy function approximation: a gradient boosting machine. 1189-1232.
Geurts, P., Ernst, D. & Wehenkel, L. 2006. Extremely randomized trees. Machine learning, **63**, 3-42.
Ho, T.K. 1995. Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 278-282.
Hough, B.K. 1957. Basic Soils Engineering.
Kalantary, F. & Kordnaeij, A. 2012. Prediction of compression index using artificial neural network. Scientific Research and Essays, **7**, 2835-2848.
Kennedy, J. & Eberhart, R. 1995. Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks, **4**, 1942-1948, doi: 10.1109/ICNN.1995.488968.
Koppula, S. 1981. Statistical estimation of compression index. ASTM Geotechnical Testing Journal, **4**.
Kumar, V.P. & Rani, C.S. 2011. Prediction of compression index of soils using artificial neural networks (ANNs). Int J Eng Res Appl, **1**, 1554-1558.
Laan, M.J.v.d., Polley, E.C. & Hubbard, A.E. 2007. Super Learner. Statistical Applications in Genetics and Molecular Biology, **6**, doi: doi:10.2202/1544-6115.1309.
LCPC. 1977. Remblais sur sols compressibles. Bulletin des Laboratoires des Ponts et Chaussees, Special T, 58. Paris (France). 361p.
Little, R.J. & Rubin, D.B. 2019. Statistical analysis with missing data. John Wiley & Sons.
Long, T., He, B., Ghorbani, A. & Khatami, S.M.H. 2023. Tree-Based Techniques for Predicting the Compression Index of Clayey Soils. Journal of Soft Computing in Civil Engineering, **7**, 52-67.
McCabe, B.A., Sheil, B.B., Long, M.M., Buggy, F.J. & Farrell, E.R. 2014. Empirical correlations for the compression index of Irish soft soils. Proceedings of the Institution of Civil Engineers-Geotechnical Engineering, **167**, 510-517.
Mitachi, T. & Ono, T. 1985. Prediction of undrained shear strength of overconsolidated clay. Tsuchi to Kiso, JSSMFE, **33**, 21-26.

Nesamatha, R. & Arumairaj, P. 2015. Numerical modeling for prediction of compression index from soil index properties. Electron J Geotech Eng, **20**, 4369-4378.
Nishida, Y. 1956. A brief note on compression index of soil. Journal of the Soil Mechanics and Foundations Division, **82**, 1027-1021-1027-1014.
Onyejekwe, S., Kang, X. & Ge, L. 2015. Assessment of empirical equations for the compression index of fine-grained soils in Missouri. Bulletin of Engineering Geology and the Environment, **74**, 705-716.
Onyejekwe, S., Kang, X. & Ge, L. 2016. Evaluation of the scale of fluctuation of geotechnical parameters by autocorrelation function and semivariogram function. Engineering Geology, **214**, 43-49.
Park, H.I. & Lee, S.R. 2011. Evaluation of the compression index of soils using an artificial neural network. Computers and Geotechnics, **38**, 472-481.
Reagan, L.A. & Kiemele, M.J. 2008. Design for six sigma: The tool guide for practitioners. CTQ Media.
Rendon-Herrero, O. 1980. Universal compression index equation. Journal of the Geotechnical Engineering Division, **106**, 1179-1200.
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.
Samui, P., Kim, D., Das, S. & Yoon, G.L. 2012. Determination of Compression Index for Marine Clay: A Relevance Vector Machine Approach. Marine Georesources & Geotechnology, **30**, 263-273, doi: 10.1080/1064119X.2011.614323.
Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J. & Platt, J. 1999. Support vector method for novelty detection. Advances in neural information processing systems, **12**.
Singh, M.J., Kaushik, A., Patnaik, G., Xu, D.-S., Feng, W.-Q., Rajput, A., Prakash, G. & Borana, L. 2023. Machine learning-based approach for predicting the consolidation characteristics of soft soil. Marine Georesources & Geotechnology, 1-15, doi: 10.1080/1064119X.2023.2193174.
Sobol, I.Y.M. 1990. On sensitivity estimation for nonlinear mathematical models. Matematicheskoe modelirovanie, **2**, 112-118.
Spagnoli, G. & Shimobe, S. 2020. Statistical analysis of some correlations between compression index and Atterberg limits. Environmental Earth Sciences, **79**, 532.
Sridharan, A. & Nagaraj, H. 2000. Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties. Canadian Geotechnical Journal, **37**, 712-722.
Terzaghi, K., Peck, R.B. & Mesri, G. 1967. Soil mechanics in engineering practice. 2nd edition. John Wiley and Sons, New York.
Trong, D.K., Pham, B.T., Jalal, F.E., Iqbal, M., Roussis, P.C., Mamou, A., Ferentinou, M., Vu, D.Q., Duc Dam, N. & Tran, Q.A. 2021. On random subspace optimization-based hybrid computing models predicting the california bearing ratio of soils. Materials, **14**, 6516.
Tsuchida, T. 1991. A new concept of e-logp relationship for clays. *Proce, 9th Asian regional Conference on soil mechanics and foundation Engineering*, 87-90.
Van Buuren, S. & Oudshoorn, C.G. 2000. Multivariate imputation by chained equations. Leiden: TNO.
Verbrugge, J.-C. & Schroeder, C. 2018. Geotechnical correlations for soils and rocks. John Wiley & Sons.
Widodo, S. & Ibrahim, A. 2012. Estimation of primary compression index (Cc) using physical properties of Pontianak soft clay. International Journal of Engineering Research and Applications, **2**, 2231-2235.
Wroth, C. & Wood, D. 1978. The correlation of index properties with some basic engineering properties of soils. Canadian Geotechnical Journal, **15**, 137-145.
Zhang, P., Yin, Z.-Y., Jin, Y.-F., Chan, T.H.T. & Gao, F.-P. 2021. Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geoscience Frontiers, **12**, 441-452, doi: https://doi.org/10.1016/j.gsf.2020.02.014.

Zhang, W., Gu, X., Hong, L., Han, L. & Wang, L. 2023. Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges. Applied Soft Computing, **136**, 110066, doi: https://doi.org/10.1016/j.asoc.2023.110066.
Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D. & Zhang, Y. 2022. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Research, **109**, 1-17.