Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Tree-Based Techniques for Predicting the Compression Index of Clayey Soils

Long Tsang[1], Biao He[2*], Ali Ghorbani[3], Seyed Mohammad Hossein Khatami[4]

1. Geofirst Pty Ltd., 2/7 Luso Drive, Unanderra, NSW 2526, Australia
2. Ph.D. Student, Department of Civil Engineering, Faculty of Engineering, Universiti Malaya, 50603, Kuala Lumpur, Malaysia
3. Assistant Professor, Department of Engineering, Payame Noor University, Tehran, Iran
4. Department of Civil Engineering, Technical and Vocational University (TVU), Tehran, Iran

Corresponding author: *s2005282@siswa.um.edu.my*

https://doi.org/10.22115/SCCE.2023.377601.1579

ARTICLE INFO

Article history:
Received: 28 December 2022
Revised: 16 January 2023
Accepted: 31 January 2023

Keywords:
Compression index;
Consolidation;
Machine learning;
Random forest;
Extreme gradient boosting.

ABSTRACT

Compression index is an effective assessment of primary consolidation settlement of clayey soils, but the process of obtaining compression index is time-consuming and laborious. Thus, in the present study, we developed two classical tree-based techniques: random forest (RF) and extreme gradient boosting (XGBoost), to predict the compression index of clayey soils. To establish these two models, we collected an available dataset—including 391 consolidation tests for soils—from previously published research. The dataset consists of six physical parameters, including the initial void ratio, natural water content, liquid limit, plastic index, specific gravity, and soil compression index. The first five parameters are the models' inputs while the compression index is the models' output. We trained both two tree-based models using 90% of the entire dataset and used the remaining 10% to assess the well-trained models, which is consistent with the published research. Several statistical metrics, such as coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are the criteria for assessing the models' performance. The results show that the RF model has better accuracy in predicting compression index compared with the XGBoost model because it outperforms the XGBoost model both on the training and testing datasets. The performance of the RF model is R^2 of 0.928 and 0.818, RMSE of 0.016 and 0.025, MAPE of 7.046% and 10.082%, and MAE of 0.012 and 0.020 on the training and testing datasets, respectively. The sensitivity analysis reveals that the initial void ratio has a significant impact on the compression index of clayey soils.

How to cite this article: Tsang L, He B, Ghorbani A, Khatami SMH. Tree-Based Techniques for Predicting the Compression Index of Clayey Soils. J Soft Comput Civ Eng 2023;7(3):52–67. https://doi.org/10.22115/scce.2023.377601.1579

1. Introduction

The compression index of clayey soils is a measure of the soil's ability to compress or consolidate under an applied load [1]. It is a crucial property for engineers to consider when designing foundations, as it can affect the stability and settlement of the structure. The compression index of clayey soils is typically determined by conducting a series of oedometer laboratory tests on soil samples [2,3]. These tests involve applying increasing levels of stress to the soil and measuring the resulting consolidation or compression. The compression index is then calculated based on the amount of compression that occurs under a given stress level [4]. In general, clayey soils with a high compression index will be more prone to settlement and instability under load, while those with a low compression index will be more stable [5,6]. Engineers must consider the compression index of the soil when designing foundations to ensure that the structure is stable and will not experience excessive settlement.

Since conducting oedometer tests is time-consuming, costly, and unwieldy, scholars tried to create empirical formulas to predict the compression index [7–10]. However, most empirical formulas are based on the on-site environment and thereby their universality is insufficient. The empirical formula may not account for variations in soil properties and conditions that can affect the compression index. Additionally, empirical models are based on a limited amount of data and may not be accurate for all types of clay soils [11–13].

Encouragingly, with the rapid development of the soft computing technique, many scholars paid attention to its computational efficacy and high accuracy. Since the soft computing technique has been successfully used in different disciplines of civil engineering [14–21], researchers attempted to apply the soft computing technique to establish the relationship between the basic soil properties and the compression index [22,23]. Kurnaz et al. developed an artificial neural network (ANN) model to predict the compression and recompression index. The model was built on a dataset that consists of 246 laboratory oedometer tests, and the model’s inputs (soil properties) included the natural water content, liquid limit, plastic index, and specific gravity of soil particles [24]. Kordnaeij et al. proposed a group method of data handling (GMDH) type neural network to predict the recompression index. The used dataset, compiled from 344 consolidation tests for soils, included the soil properties such as the liquid limit, initial void ratio, specific gravity, natural water content, plastic index, and dry density [25]. Nguyen et al. proposed a hybrid ANN model: Biogeography-Based Optimization ANN. They used 188 soil samples to build the hybrid ANN model. The input parameters include the depth of samples, clay, moisture content, bulk density, dry density, specific gravity, void ratio, porosity, degree of saturation, liquid limit, plastic limit, plastic index, and liquid index. The principle component analysis (PCA) was used to reduce the dimension of input parameters [26]. Benbouras et al. exploited the performance of the multilayer neural network, genetic programming, and multiple regression in predicting the compression index. They used 373 oedometer test samples to develop the machine learning models. The best prediction model was established based on the input variables: wet density, water content, liquid limit, plastic index, void ratio, and fine contents [27].

Overall, the above-mentioned researches mainly focus on the ANN or ANN-based models. To the best knowledge of the authors, no relevant researches discuss the application prospect of tree-based models in predicting the compression index of clayey soils. Considering the merits of tree-based models, for example, they can handle a large number of features and still maintain good accuracy, and they are easy to be interpreted and explained because they are based on a set of decision trees [28], we propose a hypothesis: the tree-based model could perform well in this topic. Based on this, we will develop the models for predicting the compression index using the tree-based technique. The developed tree-based models are random forest (RF) and extreme gradient boosting (XGBoost). First, we collected a dataset of clayey soils from a published article (Ref. [29]) to establish these two tree-based models. Meanwhile, we used the grid search algorithm to seek the optimal hyperparameters of the models. By comparing their performance using some evaluation metrics, we finally determined the best model for predicting the compression index of clayey soils. Our main contribution is: we verified the promising application of tree-based models in predicting the compression index.

The rest of the paper is organized as follows: Section 2 presents the background of the data source; Section 3 describes the principle and implementation of the tree-based models; Section 4 discusses the main results of modeling; Section 5 summarizes the main conclusions.

2. Materials

In the present study, we collected a dataset that includes 391 experimental samples from a previously published article. The dataset is composed of the experimental results of consolidation testes (ASTM D 2435-96) for soils that were sampled at 125 construction sites in the north of Iran [29]. It mainly contains the physical properties of clayey soils, such as natural water content (ω_n), liquid limit (LL), plastic index (PI), initial void ratio (e_0), the specific gravity of soil particles (G_s), and compression index (C_c). Our goal is to build an effective relationship between the compression index and another five physical properties of clayey soils, with the help of the tree-based machine learning models.

Before beginning to develop the tree-based models, we need to do pre-processing on the dataset. Since the experimental tests may be subject to human-induced error, outliers could exist in the dataset, which will harm the performance of tree-based models. Thus, we use the boxplot method to detect the outliers of the dataset—which is a common way in statistics [30]. Boxplot can show the visualization of the five-number summary: the extreme lower (Min), the extreme upper (Max), the first quartile (Q1), the third quartile (Q3), and the median. Figure 1 shows the data distribution of the physical properties of clayey soils. The box extends from Q1 to Q3 of the data; the red line and rhombus point represent the median and mean values, respectively; and the black circle point denotes the outliers of each variable [31]. Intuitively, the outliers exist in each variable and should be removed. After removing the outliers, 349 data samples were available to develop the tree-based models. Table 1 presents the statistical indices of each variable in the new dataset. We can find that the range of natural water content is between 12.7% and 42.1%; the range of liquid limit is between 24% and 62%; the range of the plastic index is between 4% and

37%; the range of initial void ratio is between 0.476 and 1.059; the range of specific gravity of soil particles is between 2.5 and 2.77; the range of compression index is between 0.05 and 0.385.

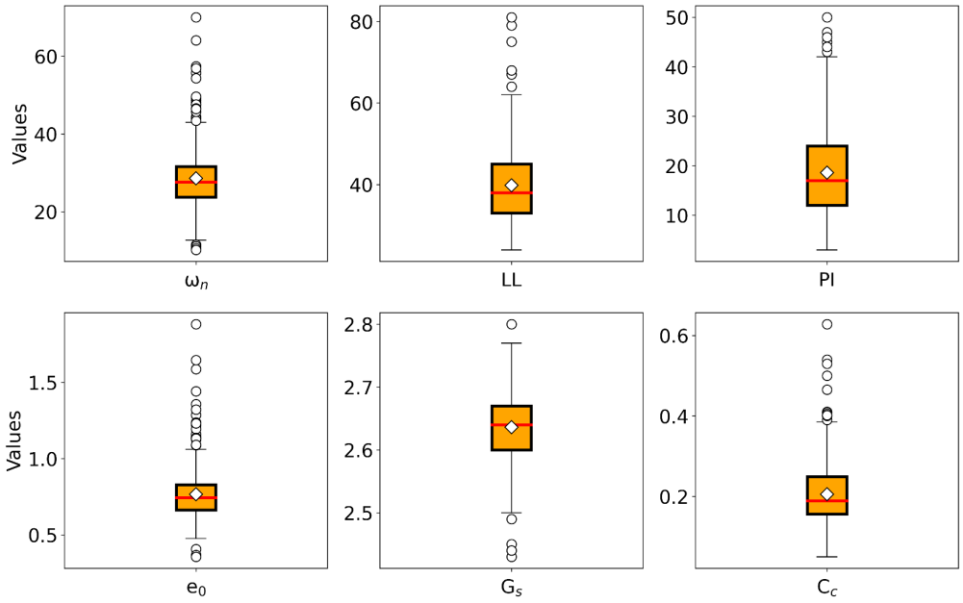

Fig. 1. Boxplot of the collected Cc dataset.

Table 1
Statistical indices of the cleaned C_c dataset.

Variables	Symbol	Unit	Min.	Max.	Mean	Std. Dev.
Natural water content	ω_n	%	12.7	42.1	27.418	5.427
Liquid limit	LL	%	24	62	38.883	8.675
Plastic index	PI	%	4	37	17.848	7.633
Initial void ratio	e_0	-	0.476	1.059	0.739	0.115
Specific gravity of soil particles	G_s	-	2.5	2.77	2.64	0.054
Compression index	C_c	-	0.05	0.385	0.194	0.059

When conducting the modeling, a common way is to divide the entire dataset into two parts: the training dataset and the testing dataset. In this way, it can effectively examine the model's generalization ability and help in avoiding overfitting. Thus, we randomly split the 349 data samples into two sets: one is the training dataset (90% of the entire data) which has 314 samples, and the other one is the testing dataset (10% of the entire data) which has 35 samples. Such a splitting strategy is consistent with the published article [29], and we anticipate verifying whether our developed models have a better performance compared with the model in that published article. We use the training dataset to establish the tree-based models for predicting the compression index and then use the testing dataset the examine the models' generalization ability. To make the random division valid, a key rule that should be obeyed is to keep the training and testing datasets have similar statistical properties. Herein, we used the cumulative

distribution function to judge whether the training dataset has acceptable statistical similarity with the testing dataset.

A cumulative distribution function can describe the probability distribution of a continuous random variable, and it is a non-decreasing function that ranges from 0 to 1 as the value of the random variable increases from negative infinity to positive infinity [32]. Figure 2 illustrates the cumulative distribution of variables in the training and testing datasets. We find that the variables: ω_n, e_0, G_s, and C_c, in both the training and testing datasets, have similar tendencies and shapes. But for variables: LL and PI, they have slight differences because the line's position of the testing dataset is below that of the training dataset. This might be because the training dataset involves more instances compared with the testing dataset, which incurs that the variables (LL and PI) have lower cumulative probabilities. Additionally, we also observe that the range of each variable in the training dataset almost covers that in the testing dataset—according to the x-axis in each subplot. This can confirm that the models fitted on the training dataset would show promising performance on the testing dataset. From the above analysis, we believe that the division of the training and testing dataset is reasonable, and they can be used to conduct the modeling accordingly.

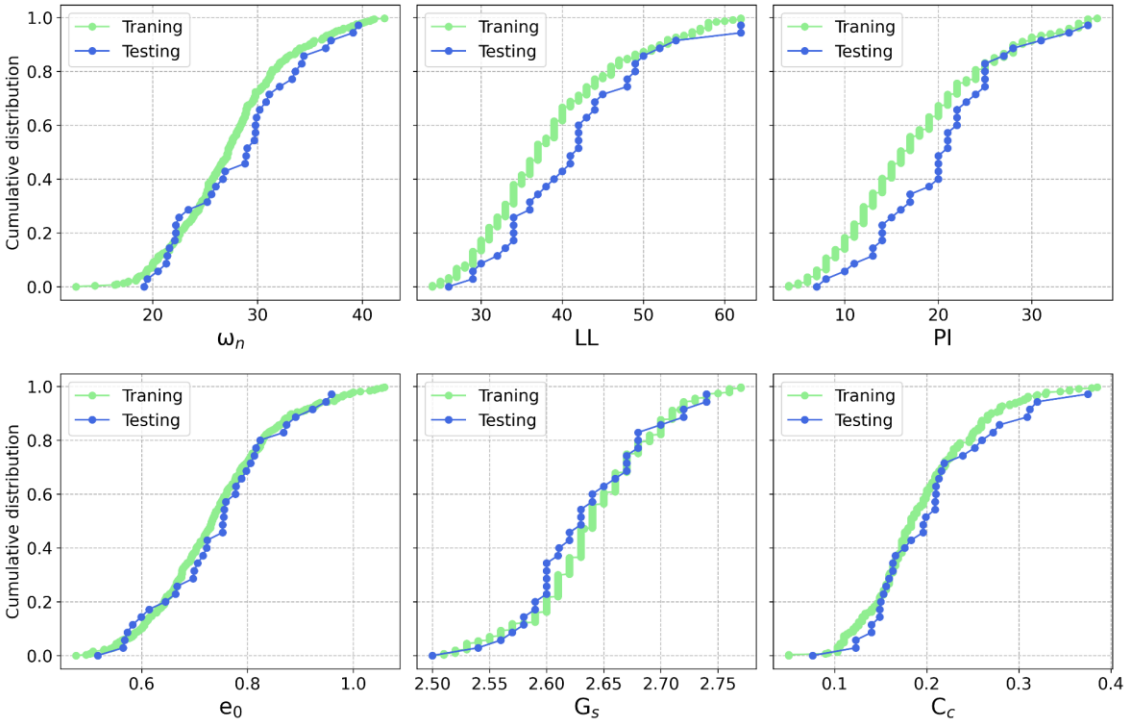

Fig. 2. Cumulative distribution of variables in training and testing datasets.

Furthermore, we present the linear relationship between the compression index and each variable—only applied to the training dataset, as shown in Figure 3. Intuitively, the initial void ratio (e_0) has a relatively strong relationship with the compression index, followed by the natural water content (ω_n). As for the liquid limit (LL), plastic index (PI), and specific gravity of soil particles (Gs), they all show an insignificant relationship with the compression index. From this point, we believe that a sophisticated model should be constructed to characterize the intrinsic

relationship between these variables and the compression index. The subsequent sections will discuss it deeply.

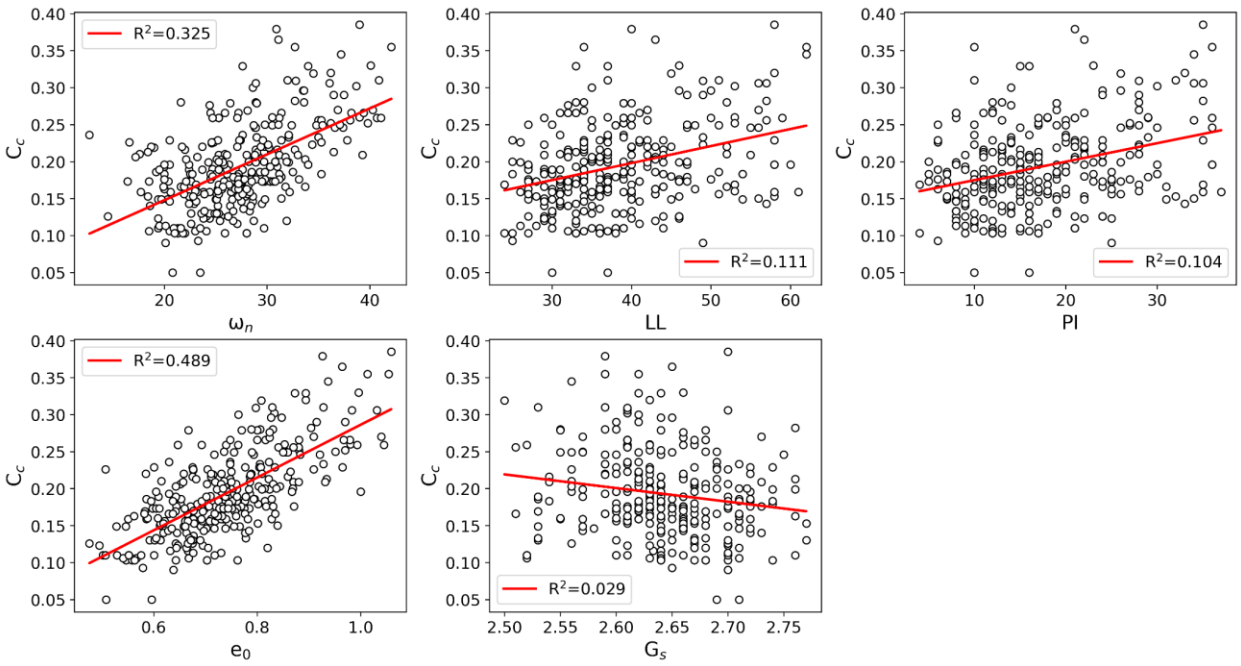

Fig. 3. Relationship between compression index and soil properties.

3. Methods

3.1. Random forest (RF)

RF is a supervised machine learning algorithm that is used for both classification and regression tasks [33]. It is an ensemble model that is composed of multiple decision trees, which are trained on different samples of the training data and then aggregated to make a final prediction. The key idea behind RF is to create a diverse set of decision trees, each of which is trained on a randomly selected subset of the training data and a randomly selected subset of the features. This process, known as bootstrapping, helps to reduce the variance of the model and make it more robust. During the training process, each decision tree in the RF makes a prediction based on the features in its training set. The final prediction of the RF is then made by aggregating the predictions of all the individual decision trees, for example, by taking the average for regression tasks, as shown below:

$$y = \frac{1}{K}\sum_{i=1}^{K} T_i(x) \tag{1}$$

where y represents the average of prediction results, K is the number of decision trees, and $T_i(x)$ represents the predicted results of a single decision tree.

One of the main benefits of using a random forest model is that it can handle large amounts of data and a high number of features, and it is also resistant to overfitting. Additionally, it is

relatively easy to interpret and understand, as the individual decision trees are simple models that can be inspected and analyzed.

RF has two key hyperparameters: the number of trees and the maximum depth, which can highly affect its performance [34]. The number of trees denoted how many decision trees are in a forest and it dramatically controls the prediction accuracy of the RF model. For another hyperparameter: maximum depth, its role is to reduce the RF model's complexity to avoid possible overfitting. In the present study, we aim to seek the optimal values of these two hyperparameters and thus construct a high-performance RF model for predicting compression index.

3.2. Extreme gradient boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a supervised machine learning algorithm that is used for both classification and regression tasks [35]. It is an ensemble model that is composed of multiple decision trees, which are trained sequentially in a way that allows the model to learn and improve from the mistakes made by earlier trees.

XGBoost is a variant of the gradient boosting algorithm, which is a type of boosting algorithm that is based on the concept of boosting weak learners to form a strong learner. Boosting algorithms work by iteratively adding weak models to the ensemble and adjusting the weights of the training data so that the mistakes made by the previous models are emphasized and corrected in the subsequent models [36]. In the XGBoost model, the decision trees are trained using gradient descent to minimize the loss function, which measures the difference between the predicted values and the true values in the training data. The loss function of the XGBoost model is shown below:

$$X_{obj} = \sum_{i=1}^{n} l\left(y, y\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{2}$$

where X_{obj} represents the objective function, $\sum_{i=1}^{n} l\left(y, y\right)$ represent the predictive loss between the predicted and real values, $\sum_{k=1}^{K} \Omega\left(f_k\right)$ represents the regularization term that is used to avoid overfitting. In general, the technique of minimizing a quadratic function is the way to optimize the objective function [37].

When constructing the XGBoost model, two key hyperparameters should be considered, that is, the number of trees and the learning rate. The number of trees refers to the maximum number of gradient-boosted trees. It controls the predictive accuracy of the XGBoost model. In general, if its value is too low/high, the model will encounter underfitting/overfitting. The learning rate refers to the step size shrinkage in each iteration. It can make the boosting process more conservative. In the present study, we aim to seek the optimal values of these two hyperparameters and thus construct a high-performance XGBoost model for predicting compression index.

3.3. Evaluation criteria

To quantitatively assess the accuracy of the above-mentioned RF and XGBoost models, some commonly used regression evaluation metrics are utilized, for example, coefficient of determination (R^2), root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The following equations are used to compute these metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i\right)^2} \tag{4}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y_i}{y_i}\right| \times 100\% \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i\right| \tag{6}$$

where y_i denotes the measured compression index, y_i denotes the predicted compression index, $\overline{y}$ denotes the average of y_i, and N is the number of samples. For the above four metrics, the closer the R^2 to 1, the better the model's performance; the smaller the RMSE, MAPE, and MAE, the better the model's performance.

3.4. Study step

The main step of the research method in the present study are as below:

As mentioned previously, the entire dataset is divided into two parts: the training dataset involving 314 soil samples and the testing dataset involving 35 soil samples. Then, we use the training dataset to establish the RF and XGBoost models, respectively. In this process, the grid search algorithm is employed to seek the optimal hyperparameters of the RF and XGBoost models [38]. The hyperparameters of the RF model are the number of trees and the maximum depth, and the hyperparameters of the XGBoost model are the number of trees and the learning rate. Meanwhile, a five-fold cross-validation algorithm is used when training the RF and XGBoost models, which aims to help the models avoid overfitting. After determining the hyperparameters of the RF and XGBoost models, we use the testing dataset to examine their generalization ability. Lastly, we also analyze which variable is highly sensitive for predicting the compression index of clayey soils. Figure 4 displays the flowchart of the present study. In the present study, we used two open-source Python libraries: Scikit-learn [39] and XGBoost [35,36] to develop the RF and XGBoost models, respectively.

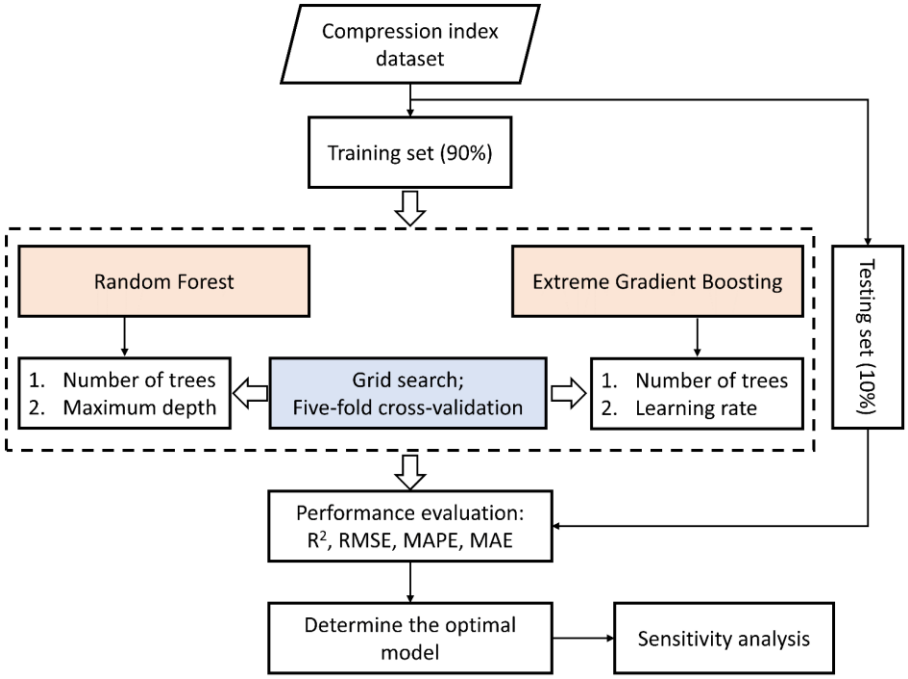

Fig. 4. Flowchart of the study step.

4. Results

4.1. Evaluation of model performance

In this section, we trained the RF and XGBoost on the same training dataset and then determined their respective optimal hyperparameters. First, we defined the searching domain of these two models, as shown in Table 2. Regarding the RF model, the searching domain of its hyper-parameters is: the number of trees increases from 50 to 300 with the increment of 10, and the maximum depth increases from 1 to 20 with the increment of 1. Regarding the XGBoost model, the searching domain of its hyperparameters is: the number of trees increases from 50 to 300 with the increment of 10, and the learning rate increases from 0.01 to 0.30 with the increment of 0.01. We then use the mean squared error (MSE) as an evaluation metric to determine the optimal hyperparameters of each model.

Figure 5 illustrates the possible results of hyperparameters of the RF model. We can find that the MSE reached a relatively large value when the maximum depth is less than 7. When the maximum depth is larger than 7, the value of MSE does not fluctuate strongly. Another point is that the maximum depth has a significant influence on the MSE compared with the number of trees because the MSE significantly reduced with the increase of the maximum depth. Conclusively, according to Figure 5 (b), the optimal hyperparameters of the RF model are: the number of trees is 130 and the maximum depth is 10.

Figure 6 illustrates the possible results of hyperparameters of the XGBoost model. We can find that the MSE reached a relatively large value only when the number of trees and learning rate are both in small values. For other cases, the MSE does not have obvious changes. According to

Figure 6 (b), we determined the optimal hyperparameters of the XGBoost model, that is, the number of trees is 60 and the learning rate is 0.14.

Table 2
Hyperparameters of the RF and XGBoost models.

Model	Hyperparameter	Searching domain	Increment
RF	Number of trees	[50, 300]	10
	Maximum depth	[1, 20]	1
XGBoost	Number of trees	[50, 300]	10
	Learning rate	[0.01, 0.30]	0.01

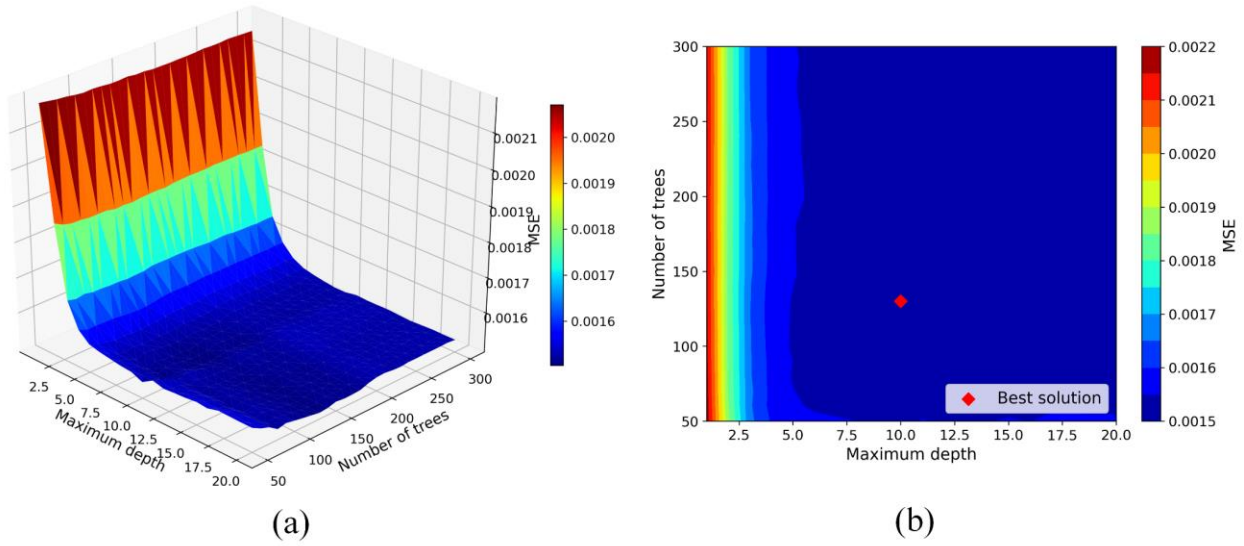

Fig. 5. Determination of the hyperparameters of the RF model.

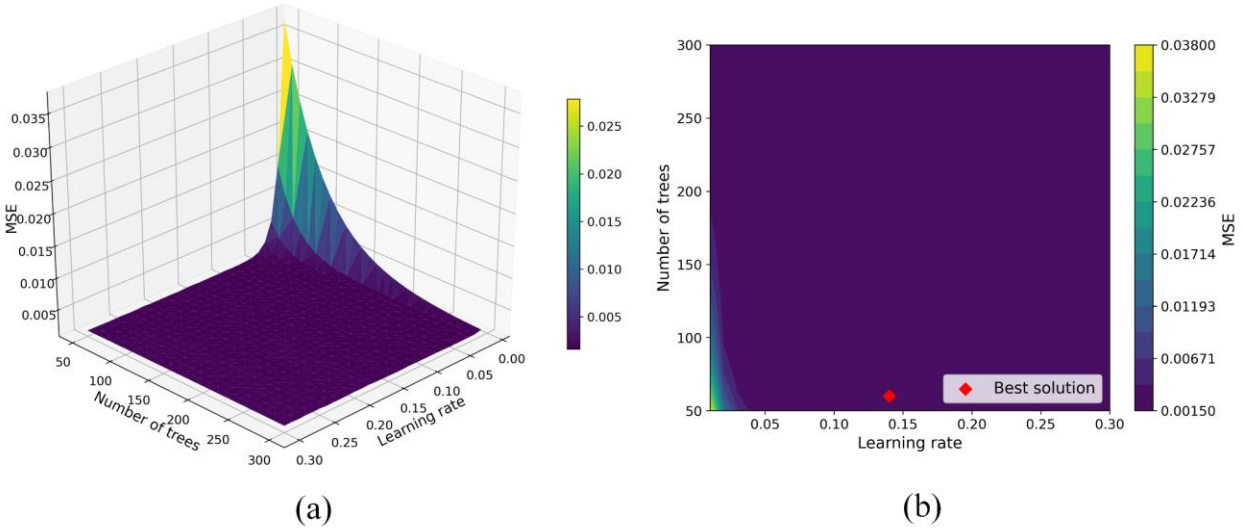

Fig. 6. Determination of the hyperparameters of the XGBoost model.

After figuring out the values of the hyperparameters of each model, we use the obtained hyperparameters to construct the RF and XGBoost models, respectively. Subsequently, we examined their performance on both training and testing datasets. At the same time, we also compared them with the model (ANN) in the published article [29]. Table 3 shows the performance of the RF and XGBoost models on the training and testing datasets. As a result, the RF model has the lowest error on both training and testing datasets compared with the XGBoost and ANN models. Its performance indices are as follows: R^2 of 0.928, RMSE of 0.016, MAPE of 7.046%, and MAE of 0.012 on the training dataset; R^2 of 0.818, RMSE of 0.025, MAPE of 10.082%, and MAE of 0.020 on the testing dataset. Additionally, we can also find that both RF and XGBoost models outperform the ANN model. Thus, we conclude that the tree-based models have a promising prospect of predicting the compression index of clayey soils.

Figure 7 shows the comparison between the experimental compression index and the predicted compression index by the RF model. Intuitively, for the training dataset, almost all the data points are concentrated around the black dashed line. This indicates the compression index predicted by the RF model approximates the experimental compression index. As for the testing dataset, most of the data points are concentrated around the black dashed line, but several data points are not. This indicates although the generalization ability of the current RF model is acceptable, it still needs further improvement. Overall, the developed RF model shows acceptable and effective performance on both the training and testing datasets.

Table 3
Comparison of models' performance.

Model	Training dataset				Testing dataset			
	R^2	RMSE	MAPE (%)	MAE	R^2	RMSE	MAPE (%)	MAE
RF	0.928	0.016	7.046	0.012	0.818	0.025	10.082	0.020
XGB	0.832	0.024	10.933	0.019	0.833	0.026	11.125	0.021
ANN [29]	-	0.035	13.340	0.027	-	0.034	13.170	0.027

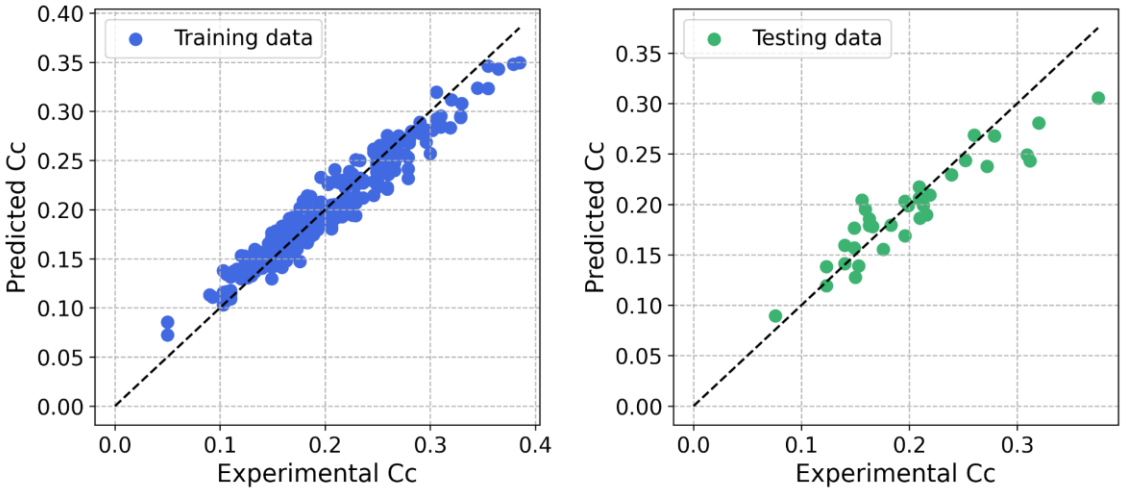

Fig. 7. Predicted and experimental compression index.

4.2. Sensitivity analysis

From the above analysis, we have successfully obtained the best tree-based model for predicting compression index, namely, the RF model. In this section, we will figure out which variable shows the highest influence on predicting compression index when using the RF model. The RF model has an intrinsic attribute: feature importance, which can measure the importance of each feature when constructing a split node in a decision tree. The standard for constructing the split node is "squared error" when the prediction is a regression task. Its main principle is to minimize the L2 loss using the mean of each split node [40]. In short, the more times the feature is used in a split node to minimize the L2 loss, the higher its importance. Based on this, we can obtain the importance of each feature (variable), as shown in Figure 8. Intuitively, for the present engineering instance, the variable e_0, i.e., the initial void ratio, shows the highest impact on predicting compression index; the variables G_s and ω_n, i.e., the specific gravity of soil particles and natural water content, show relatively slight impact on predicting compression index; the variables PI and LL, i.e., the plastic index and the liquid index, show negligible impact on prediction compression index. As a result, we conclude that the initial void ratio should be a significant concern in predicting the compression index.

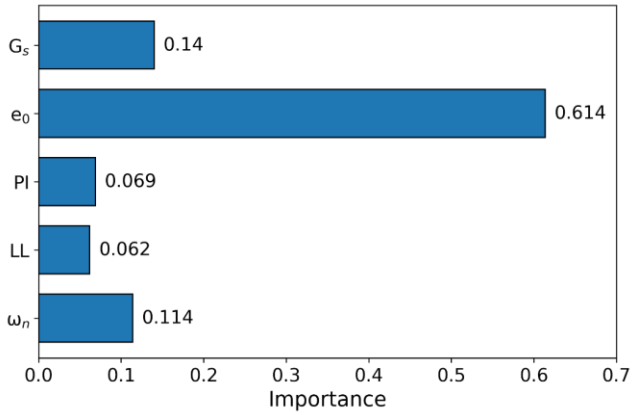

Fig. 8. Importance of each variable in predicting compression index.

Further, to identify the specific effect of the initial void ratio (e_0) on the compression index, we used the Partial Dependence Plots and Individual Conditional Expectation plots to achieve visualization and analysis of the interaction between the initial void ratio and the compression index. In general, Partial Dependence Plots can show the average (overall) dependence between the target response and the input feature of interest [41]. Individual Conditional Expectation plots can reflect the individual dependence between the target response and the input feature of interest—based on the selected data samples [42]. Figure 9 shows the particular effect of the initial void ratio on the compression index. The red dashed line represents the average dependence between the initial void ratio and the compression index. Intuitively, the compression index increases with the increase of the initial void ratio, especially when the initial void ratio is between 0.58 and 0.90. However, when the initial void ratio is between 0.476 and 0.58 as well as 0.90 and 1.059, the compression index is almost unchanged. As for the individual dependence between the initial void ratio and the compression index (all blue lines), most of the data samples present a similar trend to the red dashed line—although few of them are fluctuant.

In summary, the relationship between the initial void ratio and the compression index is approximately positive linear, which is beneficial for us to determine the compression index. Some published studies also pointed out that the compression index of clayey soils is highly dependent on the initial void ratio. For instance, Tiwari and Ajmera reported a significant linear relationship between the compression index and the initial void ratio [43]. Akbarimehe et al. also concluded a valid linear correlation between the compression index and the initial void ratio through the consolidation tests [44]. Erzin et al. developed an empirical formula based on a robust optimization model and found that the compression index is more sensitive to the initial void ratio [45].

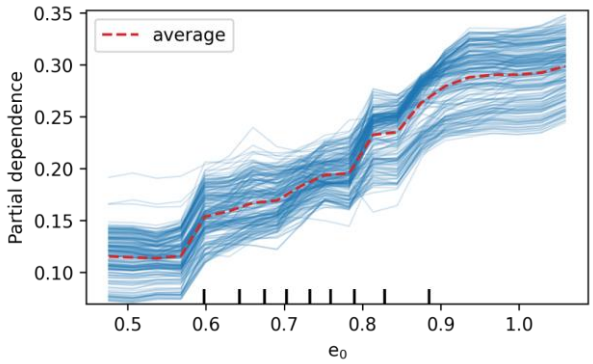

Fig. 9. Dependence between the initial void ratio and the compression index.

5. Conclusion

In the present study, we proposed two tree-based models (RF and XGBoost) to predict the compression index of clayey soils. First, a dataset for soil consolidation tests, collected from a previously published work, was utilized to develop the tree-based models. In the tree-based models, the input parameters included natural water content (ω_n), liquid limit (LL), plastic index (PI), initial void ratio (e_0), and specific gravity of soil particles (G_s), whereas the compression index is the target output. Then, we used a grid search algorithm to seek the optimal hyperparameters of the tree-based models. As a result, the optimal hyperparameters of the RF model are: number of trees = 130, maximum depth = 10, and the optimal hyperparameters of the XGBoost model are: number of trees = 60, learning rate = 0.14. By comparing their performance on both the training and testing datasets, we found that the RF model outperformed the XGBoost model. The RF model obtained the lower errors when implementing the task of predicting the compression index of clayey soils, evidenced by R^2 of 0.928 and 0.818, RMSE of 0.016 and 0.025, MAPE of 7.046% and 10.082%, and MAE of 0.012 and 0.020 on the training and testing datasets, respectively. This confirms that the RF model can help in reducing the cost of implementing laboratory experiments to determine the compression index of clayey soils. Furthermore, according to the feature importance of input parameters in the RF model, we found that the initial void ratio (e_0) has a significant impact on predicting the compression index in the present engineering instance. This is beneficial for engineers to understand the compression characteristics of clayey soils—we emphasize an approximately positive linear relationship between the initial void ratio (e_0) and the compression index of clayey soils and we recommend the engineers focus on this point when dealing with similar scenarios.

Acknowledgments

The authors would like to appreciate the Faculty of Engineering, Universiti Malaya, and the facilities provided which enabled the study to be carried out.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Authors contribution statement

L.T., B.H., A.G., S.M.H.K.: Conceptualization; L.T., B.H.: Data curation; L.T., B.H.: Formal analysis; L.T., B.H., A.G., S.M.H.K.: Investigation; L.T., B.H.: Methodology; L.T., B.H.: Software; A.G., S.M.H.K.: Supervision; A.G., S.M.H.K.: Validation; L.T., B.H.: Visualization; L.T., B.H., A.G., S.M.H.K.: Roles/Writing – original draft; L.T., B.H., A.G., S.M.H.K.: Writing – review & editing.

References

[1] Lee C, Hong SJ, Kim D, Lee W. Assessment of Compression Index of Busan and Incheon Clays with Sedimentation State. Mar Georesources Geotechnol 2015;33:23–32. https://doi.org/10.1080/1064119X.2013.764947.

[2] Nagaraj TS, Srinivasa BR, Murthy S. A critical reappraisal of compression index euqations. Geotechnique 1987:135–6.

[3] Shimobe S, Spagnoli G. A General Overview on the Correlation of Compression Index of Clays with Some Geotechnical Index Properties. Geotech Geol Eng 2022;40:311–24. https://doi.org/10.1007/s10706-021-01888-8.

[4] Onyelowe KC, Ebid AM, Nwobia L, Dao-Phuc L. Prediction and performance analysis of compression index of multiple-binder-treated soil by genetic programming approach. Nanotechnol Environ Eng 2021;6. https://doi.org/10.1007/s41204-021-00123-2.

[5] Gregory AS, Whalley WR, Watts CW, Bird NRA, Hallett PD, Whitmore AP. Calculation of the compression index and precompression stress from soil compression test data. Soil Tillage Res 2006;89:45–57. https://doi.org/10.1016/j.still.2005.06.012.

[6] McCabe BA, Sheil BB, Long MM, Buggy FJ, Farrell ER, Quigley P. Discussion: Empirical correlations for the compression index of Irish soft soils. Proc Inst Civ Eng Geotech Eng 2016;169:90–2. https://doi.org/10.1680/jgeen.15.00101.

[7] Alkroosh I, Alzabeebee S, Al-Taie AJ. Evaluation of the accuracy of commonly used empirical correlations in predicting the compression index of Iraqi fine-grained soils. Innov Infrastruct Solut 2020;5:1–10. https://doi.org/10.1007/s41062-020-00321-y.

[8] Danial Mohammadzadeh S, Kazemi SF, Mosavi A, Nasseralshariati E, Tah JHM. Prediction of compression index of fine-grained soils using a gene expression programming model. Infrastructures 2019;4:1–12. https://doi.org/10.3390/infrastructures4020026.

[9] Singh A, Noor S. Soil Compression Index Prediction Model for Fine Grained Soils. Int J Innov Eng Technol 2012;1:4.

[10] Al-khafaji AW. Compression Index Equations 2018.

[11] Yoon GL, Kim BT, Jeon SS. Empirical correlations of compression index for marine clay from regression analysis. Can Geotech J 2004;41:1213–21. https://doi.org/10.1139/t04-057.

[12] Park H Il, Lee SR. Evaluation of the compression index of soils using an artificial neural network. Comput Geotech 2011;38:472–81. https://doi.org/10.1016/j.compgeo.2011.02.011.

[13] Onyejekwe S, Kang X, Ge L. Assessment of empirical equations for the compression index of fine-grained soils in Missouri. Bull Eng Geol Environ 2015;74:705–16. https://doi.org/10.1007/s10064-014-0659-8.

[14] Ghanizadeh AR, Ghanizadeh A, Asteris PG, Fakharian P, Armaghani DJ. Developing bearing capacity model for geogrid-reinforced stone columns improved soft clay utilizing MARS-EBS hybrid method. Transp Geotech 2023;38:100906. https://doi.org/10.1016/j.trgeo.2022.100906.

[15] Cavaleri L, Barkhordari MS, Repapis CC, Armaghani DJ, Ulrikh DV, Asteris PG. Convolution-based ensemble learning algorithms to estimate the bond strength of the corroded reinforced concrete. Constr Build Mater 2022;359:129504.

[16] Tan WY, Lai SH, Teo FY, Armaghani DJ, Pavitra K, El-Shafie A. Three Steps towards Better Forecasting for Streamflow Deep Learning. Appl Sci 2022;12. https://doi.org/10.3390/app122412567.

[17] Shan F, He X, Jahed Armaghani D, Zhang P, Sheng D. Success and challenges in predicting TBM penetration rate using recurrent neural networks. Tunn Undergr Sp Technol 2022;130:104728. https://doi.org/10.1016/j.tust.2022.104728.

[18] Skentou AD, Bardhan A, Mamou A, Lemonis ME, Kumar G, Samui P, et al. Closed-Form Equation for Estimating Unconfined Compressive Strength of Granite from Three Non-destructive Tests Using Soft Computing Models. Rock Mech Rock Eng 2022:https://doi.org/10.1007/s00603-022-03046-9.

[19] Indraratna B, Armaghani DJ, Correia AG, Hunt H, Ngo T. Prediction of resilient modulus of ballast under cyclic loading using machine learning techniques. Transp Geotech 2022:100895.

[20] Ghanizadeh AR, Delaram A, Fakharian P, Armaghani DJ. Developing Predictive Models of Collapse Settlement and Coefficient of Stress Release of Sandy-Gravel Soil via Evolutionary Polynomial Regression. Appl Sci 2022;12:9986. https://doi.org/10.3390/app12199986.

[21] He B, Armaghani DJ, Lai SH. Assessment of tunnel blasting-induced overbreak: A novel metaheuristic-based random forest approach. Tunn Undergr Sp Technol 2023;133:104979. https://doi.org/10.1016/j.tust.2022.104979.

[22] Mittal M, Satapathy SC, Pal V, Agarwal B, Goyal LM, Parwekar P. Prediction of coefficient of consolidation in soil using machine learning techniques. Microprocess Microsyst 2021;82:103830. https://doi.org/10.1016/j.micpro.2021.103830.

[23] Ozer M, Isik NS, Orhan M. Statistical and neural network assessment of the compression index of clay-bearing soils. Bull Eng Geol Environ 2008;67:537–45. https://doi.org/10.1007/s10064-008-0168-8.

[24] Kurnaz TF, Dagdeviren U, Yildiz M, Ozkan O. Prediction of compressibility parameters of the soils using artificial neural network. Springerplus 2016;5:1801. https://doi.org/10.1186/s40064-016-3494-5.

[25] Kordnaeij A, Kalantary F, Kordtabar B, Mola-Abasi H. Prediction of recompression index using GMDH-type neural network based on geotechnical soil properties. Soils Found 2015;55:1335–45. https://doi.org/10.1016/j.sandf.2015.10.001.

[26] Nguyen MD, Pham BT, Ho LS, Ly HB, Le TT, Qi C, et al. Soft-computing techniques for prediction of soils consolidation coefficient. Catena 2020;195. https://doi.org/10.1016/j.catena.2020.104802.

[27] Benbouras MA, Kettab Mitiche R, Zedira H, Petrisor AI, Mezouar N, Debiche F. A new approach to predict the compression index using artificial intelligence methods. Mar Georesources Geotechnol 2019;37:704–20. https://doi.org/10.1080/1064119X.2018.1484533.

[28] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67. https://doi.org/10.1007/s10462-020-09896-5.

[29] Farzin Kalantary. Prediction of compression index using artificial neural network. Sci Res Essays 2012;7. https://doi.org/10.5897/SRE12.297.

[30] Smiti A. A critical overview of outlier detection methods. Comput Sci Rev 2020;38:100306. https://doi.org/10.1016/j.cosrev.2020.100306.

[31] Wickham H, Stryjewski L. 40 Years of Boxplots. HadCoNz 2011:1–17.

[32] Drew JH, Glen AG, Leemis LM. Computing the cumulative distribution function of the Kolmogorov-Smirnov statistic. Comput Stat Data Anal 2000;34:1–15. https://doi.org/10.1016/S0167-9473(99)00069-9.

[33] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[34] He B, Lai SH, Mohammed AS, Muayad M, Sabri S, Ulrikh DV. Estimation of Blast-Induced Peak Particle Velocity through the Improved Weighted Random Forest Technique. Appl Sci 2022;12:5019. https://doi.org/https://doi.org/10.3390/app12105019.

[35] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., New York, NY, USA: ACM; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

[36] Chen T, Guestrin C. XGBoost: Reliable Large-scale Tree Boosting System Tianqi. Proc. 22nd SIGKDD Conf. Knowl. Discov. Data Min., San Francisco, CA, USA: 2015, p. 13–7.

[37] Zhou J, Qiu Y, Zhu S, Armaghani DJ, Khandelwal M, Mohamad ET. Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization. Undergr Sp 2021;6:506–15. https://doi.org/10.1016/j.undsp.2020.05.008.

[38] Liashchynskyi P, Liashchynskyi P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS 2019:1–11.

[39] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project 2013:1–15.

[40] Biau G, Scornet E. A random forest guided tour. Test 2016;25:197–227. https://doi.org/10.1007/s11749-016-0481-7.

[41] Hastie T et. all. Statistics The Elements of Statistical Learning. Springer Ser Stat 2009;27:745.

[42] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. J Comput Graph Stat 2015;24:44–65. https://doi.org/10.1080/10618600.2014.907095.

[43] Tiwari B, Ajmera B. New Correlation Equations for Compression Index of Remolded Clays. J Geotech Geoenvironmental Eng 2012;138:757–62. https://doi.org/10.1061/(asce)gt.1943-5606.0000639.

[44] Akbarimehr D, Eslami A, Imam R. Correlations between Compression Index and Index Properties of Undisturbed and Disturbed Tehran clay. Geotech Geol Eng 2021;39:5387–93. https://doi.org/10.1007/s10706-021-01821-z.

[45] Erzin Y, MolaAbasi H, Kordnaeij A, Erzin S. Prediction of Compression Index of Saturated Clays Using Robust Optimization Model. J Soft Comput Civ Eng 2020;4:1–16. https://doi.org/10.22115/scce.2020.233075.1226.