Proceedings of the XVIII ECSMGE 2024
GEOTECHNICAL ENGINEERING CHALLENGES
TO MEET CURRENT AND EMERGING NEEDS OF SOCIETY

ISBN 978-1-032-54816-6
DOI 10.1201/9781003431749-291

ECSMGE 24
XVIII EUROPEAN CONFERENCE ON SOIL MECHANICS AND GEOTECHNICAL ENGINEERING

Estimation of permeability coefficient using machine learning algorithms on the example of AdaBoost and Artificial Neural Network algorithms

Estimation du coefficient de perméabilité à l'aide d'algorithmes d'apprentissage automatique sur l'exemple des algorithmes AdaBoost et Artificial Neural Network

J. Dzięcioł*, W. Sas
Warsaw University of Life Sciences – WULS, Warsaw, Poland
*justyna_dzieciol@sggw.edu.pl

ABSTRACT: The development of the Construction sector in recent decades and the systematic reduction of natural aggregate (NA) resources oblige the search for alternatives and substitutes for these important construction materials. The use of anthropogenic aggregates as replacements for natural aggregates provides an opportunity to reduce the number of landfills and recycle waste materials from construction and demolition (CDW), which can be successfully used in Civil Engineering. It may also involve risks that undoubtedly include estimating geotechnical parameters based on solutions developed to date for natural aggregates. A major factor affecting these estimates is that anthropogenic aggregates could have different properties and chemical compositions. This study analyzed the possibility of estimating the coefficient of permeability for Concrete Aggregate (RCA) using Machine Learning algorithms - AdaBoost and Artificial Neural Networks. The obtained results were compared with each other and with the previously used empirical formulas for estimating the coefficient of permeability in non-cohesive materials. Based on the results received, an attempt was made to interpret the physical characteristics of the aggregate that have a significant impact on the estimation of the model by the Machine Learning algorithm. The analysis carried out will allow the formulation of conclusions for further application of Machine Learning algorithms as a tool to support the prediction of the coefficient of permeability.

RÉSUMÉ: Le développement du secteur de la construction au cours des dernières décennies et la réduction systématique des ressources en granulats naturels obligent à rechercher des alternatives et des substituts à ces importants matériaux de construction. L'utilisation de granulats anthropogéniques pour remplacer les granulats naturels permet de réduire le nombre de décharges et de recycler les déchets de construction et de démolition (CDW), qui peuvent être utilisés avec succès en génie civil. Elle peut également comporter des risques, notamment en ce qui concerne l'estimation des paramètres géotechniques sur la base des solutions développées à ce jour pour les granulats naturels. L'un des principaux facteurs influençant ces estimations est le fait que les agrégats anthropogéniques peuvent avoir des propriétés et des compositions chimiques différentes. Cette étude a analysé la possibilité d'estimer le coefficient de perméabilité pour les granulats de béton (RCA) en utilisant des algorithmes d'apprentissage automatique - AdaBoost et les réseaux neuronaux artificiels. Les résultats obtenus ont été comparés entre eux et avec les formules empiriques précédemment utilisées pour estimer le coefficient de perméabilité des matériaux non cohésifs. Sur la base des résultats obtenus, on a tenté d'interpréter les caractéristiques physiques de l'agrégat qui ont un impact significatif sur l'estimation du modèle par l'algorithme d'apprentissage automatique. L'analyse effectuée permettra de formuler des conclusions pour l'application ultérieure des algorithmes d'apprentissage automatique en tant qu'outil d'aide à la prédiction du coefficient de perméabilité.

Keywords: Permeability coefficient; machine learning; AdaBoost; artificial neural network.

1 INTRODUCTION

Sustainability is the improvement of the quality of life in a healthy environment, positively affecting social, economic and environmental conditions for present and future generations. Growing awareness of environmental degradation caused by human activity is motivating stricter regulations and innovative recycling methods (Kumar Mehta, 2001; McNeil & Kang, 2013). Life cycle analysis of a material begins with its potential reuse, which shapes the sustainable nature of the material. This includes stages such as raw material acquisition, processing, use, preservation and disposal or recycling. Sustainable manufacturing seeks to reduce waste through a closed-loop economy (Cassiani et al., 2021; Vieira et al., 2016).

An example is the recycling of concrete from demolition, which becomes a valuable concrete aggregate that replaces natural aggregates in many applications, including road construction.

The use of recycled concrete aggregates in construction is a developing technique. According to Buck (Buck, 1973), after World War II, when the demolition of buildings and roads generated large amounts of waste and Europe needed to be rebuilt, the use of these aggregates began to be widely used. The process of obtaining recycled concrete aggregates involves crushing concrete, excluding rebar, bricks and soft materials (wood, glass, polystyrene), and obtaining material with fractions from 0 to 63 mm (Oikonomou, 2005). Physical properties of recycled concrete aggregates include density, moisture content, porosity, water absorption, grain shape, fractional composition and filtration properties. In road construction and dam design, the filtration coefficient is crucial. For cost reduction and waste management, the filtration coefficient can be predicted based on physical properties instead of performing costly field or laboratory tests. Machine Learning is employed to more effectively amalgamate concealed details within datasets. The objective of machine learning involves utilizing algorithms to unveil connections between attributes present in datasets, and these revelations can subsequently enhance the procedure of making informed decisions. The purpose of this publication is to demonstrate the predictive capabilities of AdaBoost and Artificial Neural Network algorithms for estimating the filtration coefficient for recycled concrete aggregates (RCA).

2 MATERIAL

The concrete aggregate used in the study came from the demolition work of a repaired road surface. The grain size curves are presented below in Table 1.

Table 1. A summary of the material's grain size range.

	Minimum	Maximum
Particle size (d_5), [mm]	0.15	0.17
Particle size (d_{10}), [mm]	0.28	0.30
Particle size (d_{17}), [mm]	0.60	0.80
Particle size (d_{20}), [mm]	0.98	1.30
Particle size (d_{30}), [mm]	1.90	2.20
Particle size (d_{50}), [mm]	3.30	3.90
Particle size (d_{60}), [mm]	5.50	7.80
Particle size (d_{90}), [mm]	9.80	12.50

Figure 1 presents the results of the coefficient of permeability tests with a compaction energy of 0.59 J/cm^3 and 2.65 J/cm^3.

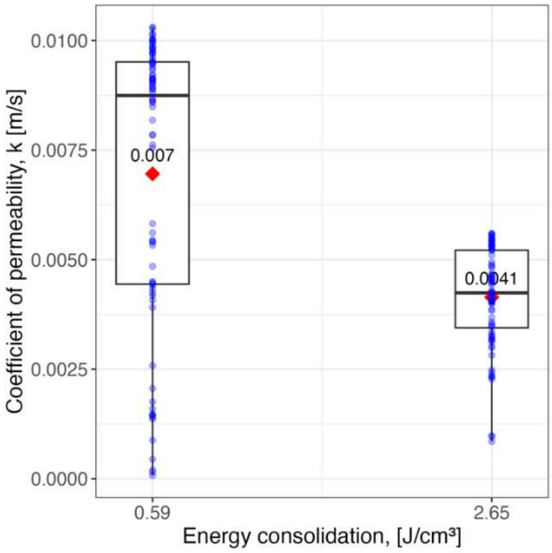

Figure 1. The results of the coefficient of permeability tests.

3 METHOD

A constant head method was applied to test the coefficient of permeability. Artificial Neural Networks and the newer boosting method, AdaBoost, were used for testing and predictive analysis. A brief overview of both Machine Learning algorithms is presented below.

Artificial Neural Networks, often referred to as ANNs or neural networks, represent a widely adopted technique in the realm of machine learning, employed to scrutinize and comprehend data by utilizing an intricate network of decision-making layers. These networks are essentially composed of interconnected decision-making entities, known as nodes, which collaborate through connections akin to the synapses of biological neurons. The arrangement of these nodes takes the form of layers, with the initial layer usually being fed with a comprehensive dataset. This initial layer constitutes raw input data, encompassing a spectrum of information formats such as numerical values, textual data, and even visual or auditory components like image pixels or sound waveforms. In the functioning of an ANN, each input node functions as a transmitter of information, conveying its input to the subsequent layer of nodes via the network's interconnections. This process enables the network to progressively synthesize and process data, capturing intricate relationships among diverse attributes (Lagaros, 2023; Shahin et al., 2021).

The core objective of AdaBoost is the amalgamation of numerous weak learning attributes into a singular potent learning attribute. This technique centers its attention on these weak learning attributes, often taking the form of decision trees with minimal partitions, akin to what is colloquially known as decision trunks. The primary emphasis lies in enhancing the performance of these less complex learning components and fusing them into a formidable collective force, capable of achieving more robust and accurate outcomes (Lin et al., 2006; Zhou & Yu, 2009).

Cross-validation serves as a fundamental tool within the realm of Machine Learning, aiming to assess a model's capacity for predicting data outcomes. The process involves a parameter known as "k," which signifies the number of partitions in which the dataset is to be segregated. Upon selecting a specific value for "k," it can be directly utilized to characterize the model; for instance, when k equals 10, it transforms into a 10-fold cross-validation (Browne, 2000).

In practical terms, cross-validation entails utilizing a limited dataset to gauge how proficiently the model is likely to perform when making predictions on new, previously unseen data. This practice is particularly advantageous as it mitigates the risk of overestimating the model's predictive prowess, which can occur when using other evaluation techniques (Fushiki, 2011).

This methodology is especially appealing due to its simplicity and comprehensibility. It contributes to the creation of a more dependable model, fostered by the learning and testing phases of its development. The cumulative outcome of k-fold cross-validation is distilled through the computation of the average skill scores exhibited by the model. To gain a comprehensive understanding, it is advisable to incorporate a measure of the fluctuation within these skill scores, which can be conveyed by parameters such as the standard deviation or standard error. The results were verified with the use of error analysis, for individual models were estimated:

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad (1)$$

- Mean Squared Residuals (MSR):

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

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_t^P| \qquad (3)$$

- Coefficient of determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad (4)$$

SHAP analysis was also performed.

4 RESULTS

4.1 Analysis of prediction results

Table 2 presents the prediction results of the permeability coefficient for both algorithms.

The following physical characteristics of the investigated Recycled Concrete Aggregate (RCA) were used to determine models based on AdaBoost and Neural Network algorithms: grain sizes at selected percentage content, including smaller sizes (d_5, d_{10}, d_{17}, d_{20}, d_{30}, d_{50}, d_{60}, d_{90}), grain size curvature index (C_C), and uniformity coefficient (C_U), volumetric density of the skeletal gravel structure, porosity index, and degree of compaction. In addition to material properties, the prediction also included the gradient at which the test was performed. It was decided to include the gradient as an important parameter of the study.

In the case of the training sample, better results of fitting to the observation data were obtained for the AdaBoost algorithm, in its case the R^2 was 0.959, for Neural Network the R^2 was 0.682.

Table 2. Evaluation of training set.

Model	RMSE	MSR	MAE	R^2
AdaBoost	0.001	3.04×10^{-7}	3.62×10^{-4}	0.959
Neural Network	0.002	2.39×10^{-6}	1.26×10^{-3}	0.682

In the test sample (Table 3) the R^2 results obtained for the algorithms were lower for AdaBoost the coefficient of determination was 0.886, and for Neural Network it was 0.640.

Table 2. Evaluation of test set.

Model	RMSE	MSR	MAE	R^2
AdaBoost	0.001	8.07×10^{-7}	6.34×10^{-4}	0.886
Neural Network	0.002	2.53×10^{-6}	1.30×10^{-3}	0.640

The results of the prediction of the filtration coefficient versus the results from the observations are presented in the figures below. For the AdaBoost algorithm for the training sample in Figure 2, for the test sample in Figure 4, the Neural Network algorithm in Figure 3 for the training sample and Figure 5 for the test sample. The R^2 results for both algorithms translate into a distance of the individual results from the reference line, and thus larger values of the residuals. Comparing all the graphs, one can notice a greater concentration of results around the reference line in the case of the AdaBoost algorithm, which harmonizes with the obtained high R^2 results and low error results for this algorithm.

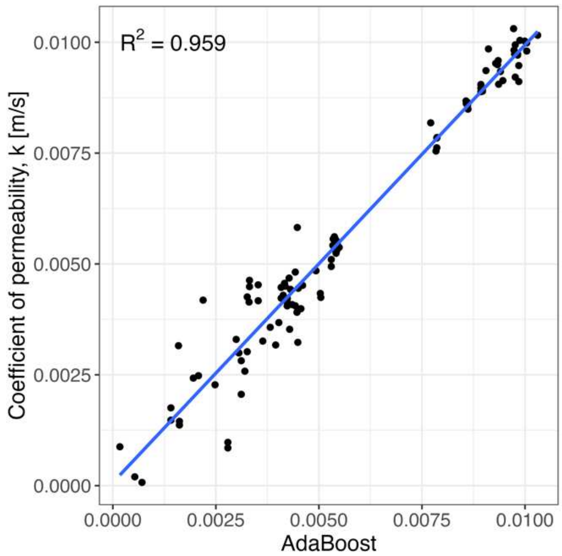

Figure 2. Training set – AdaBoost.

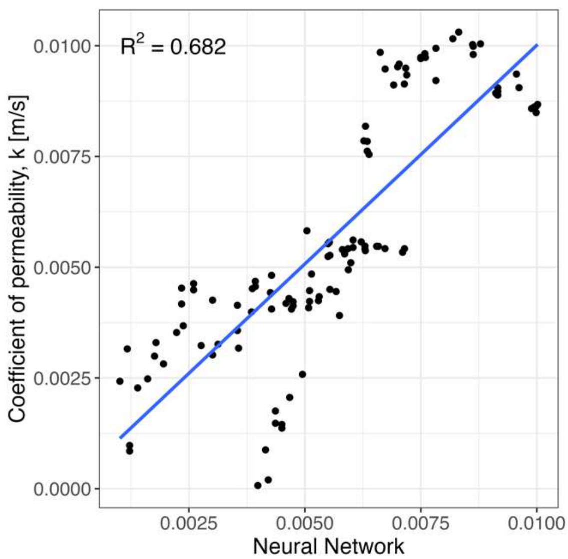

Figure 3. Training set – Neural Network.

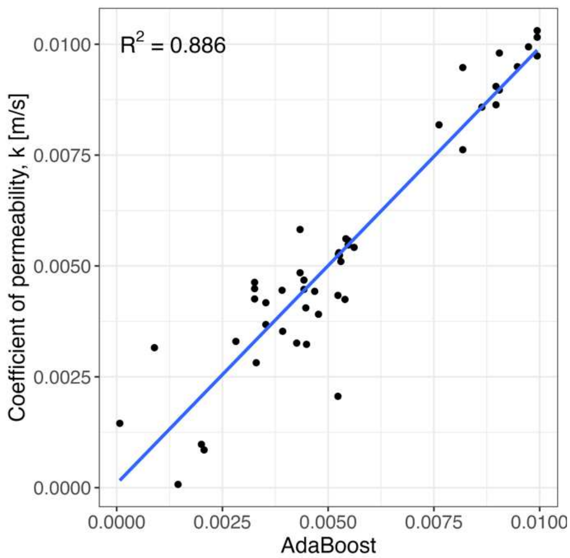

Figure 4. Test set – AdaBoost.

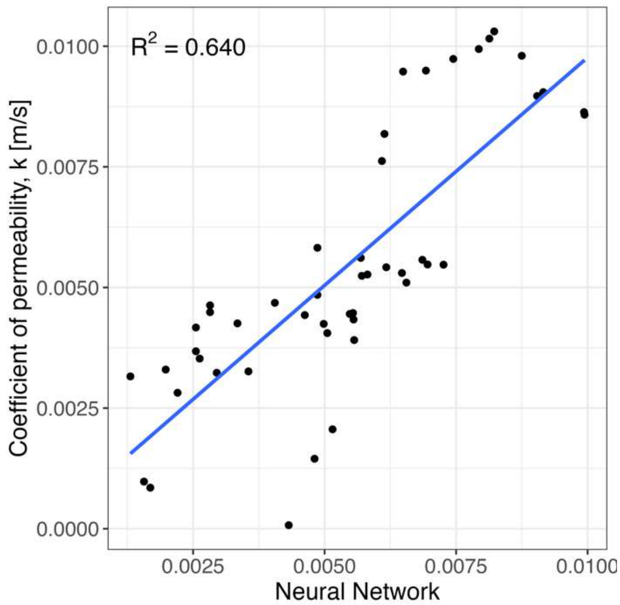

Figure 5. Test set – Neural Network.

Significant discrepancies in the results obtained by the various algorithms should be attributed to their different operating model. Adaboost and Artificial Neural Networks are two distinct approaches to machine learning, each with its characteristics and applications.

Adaboost is an ensemble learning method, specifically a boosting algorithm that combines multiple weak learning algorithms to create a strong learning algorithm. The algorithm sequentially matches these weak learning algorithms to the data, with each successive learning algorithm paying more attention to instances misclassified by previous ones. Artificial Neural Networks are a type of machine learning model inspired by the structure and functioning of the human brain. The process of learning an algorithm is usually done through backpropagation, in which errors are propagated backward through the network, updating the weights to minimize the error. Neural networks require significant amounts of labeled data to train, especially for tasks involving high-dimensionality inputs. Increasing the number of analyzed results may lead to reduced errors.

4.2 Interpretation of results using the SHAP technique

Explanatory methods for predictive models generated by machine learning algorithms offer crucial insights into the inner workings of these algorithms, enabling comprehension of the relationships among dataset features and their impact on model outcomes. Efforts to combine prediction and interpretation models give rise to an unavoidable trade-off between prediction accuracy and the ease of model interpretation. To address this, the concept of explainable artificial intelligence (XAI) comes into play. In XAI models, such as local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP), prediction and interpretation models are developed independently. These models enhance interpretability without compromising prediction accuracy.

Interpreting the decisions made by machine learning models like Artificial Neural Networks and Adaboost can be challenging due to their complex structures and opaque decision-making processes. The SHAP technique assigns specific scores to input features in the dataset based on their relevance in predicting the target value. Additionally, it is commonly applied for feature selection and dimensionality reduction, enhancing machine learning model performance by removing less pertinent features.

Applied to ANNs, the SHAP value method helps reveal which input features have the greatest impact on the network's predictions. By studying SHAP values for different features across multiple instances, patterns can be identified and insights can be gained into how the network processes information.

SHAP values can be applied to Adaboost by considering the contribution of each weak learner to the ensemble prediction for a particular instance. This involves evaluating how a change in input features affects the predictions made by each weak learner and aggregating these effects to calculate SHAP values for the ensemble's predictions. Unlike ANNs, Adaboost models are more interpretable because they consist of a collection of simpler models. Therefore, interpreting the contribution of individual weak learner models using SHAP values can provide simpler insights into model behavior.

Figures 6 and 7 in this paper employ SHAP to elucidate the role of features in prediction models, with SHAP calculating the impact of each feature on the predictions made by the learned model.

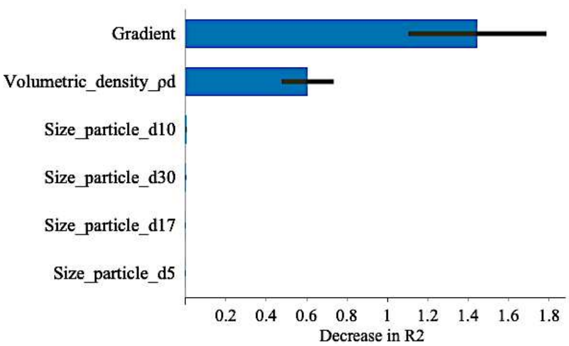

Figure 6. Feature Importance – AdaBoost.

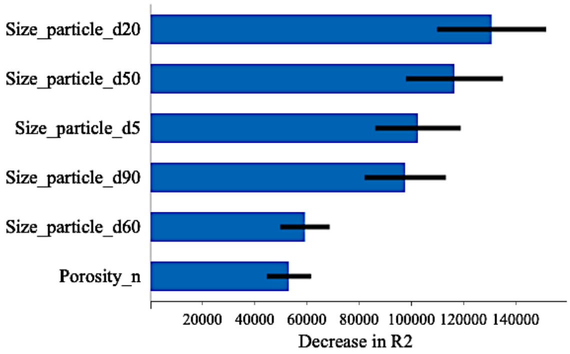

Figure 7. Feature Importance – Neural Network.

The RCA features that turned out to be particularly relevant to the model determined using the AdaBoost algorithm were gradient and volumetric density, in the second place the model took into account particle size in the case of the Neural Network where particle size d_{20}, d_{50} and d_5.

5 CONCLUSIONS

This study explores sustainable practices in recycling concrete aggregates, particularly in predicting the filtration coefficient for recycled concrete. Using machine learning techniques, AdaBoost and Neural Network algorithms were employed and compared. AdaBoost outperformed in the training sample (R^2 = 0.959), and testing sample (R^2=886). The SHAP technique enhanced interpretability, highlighting key features like gradient and volumetric density for AdaBoost and particle size (d_{20}, d_{50}, and d_5) for Neural Network. Overall, the study emphasizes the potential of machine learning in sustainable material practices.

REFERENCES

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/JMPS.1999.1279.

Buck, A. D. (1973). Recycled Concrete. *Highway Research Report*, 1–8. https://onlinepubs.trb.org/Onlinepubs/hrr/1973/430/430-001.pdf.

Cassiani, J., Martinez-Arguelles, G., Peñabaena-Niebles, R., Keßler, S., & Dugarte, M. (2021). Sustainable concrete formulations to mitigate Alkali-Silica reaction in recycled concrete aggregates (RCA) for concrete infrastructure. *Construction and Building Materials*, *307*, 124919. https://doi.org/10.1016/J.CONBUILDMAT.2021.124919.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/S11222-009-9153-8/METRICS.

Kumar Mehta, P. (2001). Reducing the Environmental Impact of Concrete. *Concrete International*, *23*(10), 61–66.

Lagaros, N. D. (2023). Artificial Neural Networks Applied in Civil Engineering. *Applied Sciences 2023, Vol. 13, Page 1131*, *13*(2), 1131. https://doi.org/10.3390/APP13021131.

Lin, G., Feng, G., Xiaohong, G., Dianmin, Z., & Jie, L. (2006). A regression algorithm based on AdaBoost. *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, *1*, 4400–4404. https://doi.org/10.1109/WCICA.2006.1713209.

McNeil, K., & Kang, T. H. K. (2013). Recycled Concrete Aggregates: A Review. *International Journal of Concrete Structures and Materials*, *7*(1), 61–69. https://doi.org/10.1007/S40069-013-0032-5/FIGURES/6.

Oikonomou, N. D. (2005). Recycled concrete aggregates. *Cement and Concrete Composites*, *27*(2), 315–318. https://doi.org/10.1016/j.cemconcomp.2004.02.020.

Shahin, M. A., Jaksa, M. B., & Maier, H. R. (2021). Artificial neural network applications in geotechnical engineering. *Australian Geomechanics*, *36*(1), 49–62.

Vieira, D. R., Calmon, J. L., & Coelho, F. Z. (2016). Life cycle assessment (LCA) applied to the manufacturing of common and ecological concrete: A review. *Construction and Building Materials*, *124*, 656–666. https://doi.org/10.1016/J.CONBUILDMAT.2016.07.125.

Zhou, Z.-H., & Yu, Y. (2009). AdaBoost. *The Top Ten Algorithms in Data Mining*, 141–164.