Acta Sci. Pol.
Architectura 22 (2023), 184–191
aspa.sggw.edu.pl ISSN 1644-0633 (suspended) eISSN 2544-1760 DOI: 10.22630/ASPA.2023.22.18

ORIGINAL PAPER

Received: 12.09.2023
Accepted: 08.12.2023

MACHINE LEARNING IN CIVIL ENGINEERING ON THE EXAMPLE OF PREDICTION OF THE COEFFICIENT OF PERMEABILITY

Justyna Dzięcioł✉

Institute of Civil Engineering, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

ABSTRACT

This paper investigates the application of the machine learning techniques in the civil engineering, focusing on the prediction of permeability coefficient. Permeability coefficient is an important parameter in various civil engineering projects including groundwater flow analysis, soil stabilisation and geotechnical engineering. Traditional methods for estimating permeability are time-consuming and often based on laboratory tests. The machine learning offers a promising approach to predict it more efficiently and accurately. This paper studies several machine-learning techniques, verifying their applicability to predict the permeability coefficient for sands. The article analysed the predictive performance of the artificial neural network (ANN), the random forest (RF), the gradient boosting (GB) and the linear regression (LR). The most accurate algorithm in this case turned out to be the gradient boosting for which the coefficient of determination was 0.995, the mean absolute error was less than 0.001 and the root mean square error was 0.001.

Keywords: machine learning, coefficient of permeability, prediction

INTRODUCTION

In the discipline of civil engineering, the integration of the machine learning has proven to be an increasingly widely developed and tested technique for improving various aspects of project planning and execution. The prediction of permeability coefficient, an important parameter in assessing fluid flow through natural and anthropogenic soils, is an example of the application of the machine learning techniques. Traditionally, an estimation of the permeability coefficient requires time- and resource-consuming laboratory tests. Machine learning offers an approach based on data from previous observations, which after looking for correlations in the data of features describing the phenomenon, allows to increase the accuracy of its estimation (Naranjo-Pérez, Infantes, Fernando Jiménez-Alonso & Sáez, 2020; Pardalos, Panos, Rassia & Tsokas, 2022). The search for new solutions to optimise investment costs is particularly important, given the changing economic situation in the construction market, which has been caused by COVID-19 and the war in Ukraine (Szymanek, 2022).

Machine learning algorithms such as regression models, decision trees and neural networks can be trained on historical data covering various soil properties and permeability values. Through a process of learning from these data, these algorithms notice relationships and patterns that are often not apparent in conventional analysis methods. The predictive power of these models lies in their ability to generalise from the data they have learned. This allows them to predict new data with a high degree of accuracy in estimating the permeability coefficient for different soil types.

Justyna Dzięcioł https://orcid.org/0000-0002-2436-9748
✉justyna_dzieciol@sggw.edu.pl

The machine learning models can uncover non-linear correlations and relationships that may be overlooked by traditional methods. This holistic understanding of complex interactions enables engineers to make more informed decisions when designing civil engineering projects where the permeability coefficient is a key consideration. Using the machine learning to predict the permeability coefficient not only exemplifies the symbiosis of technology and civil engineering but also highlights the potential for transformative advances in the broader field. As data collection and computational capabilities continue to evolve, it is foreseeable that such applications will continue to redefine traditional practices, enabling more efficient, accurate and innovative approaches to civil engineering projects (Reich, 1997; Melhem & Nagaraja, 2007; Kosinov, Trach & Trach, 2023).

Additionally, the incorporation of the machine learning to predict not only permeability coefficient, but also other geotechnical parameters is causing a paradigm shift in how civil engineers approach complex challenges. By adopting data-driven methodologies, engineers are empowered to make evidence-based decisions using insights derived from massive data sets accumulated over time. One remarkable advantage is the ability of the machine learning models to adapt to changing scenarios. As new data become available, models can be tuned and updated, ensuring that predictions remain relevant and accurate. This dynamic aspect of the machine learning fits perfectly with the dynamic nature of civil engineering projects, where conditions and variables are subject to change.

Moreover, the integration of the machine learning does not replace traditional engineering knowledge, but complements and extends it. The ability to interpret these models allows engineers to gain deeper insights into the factors affecting permeability. As the field of civil engineering evolves, the machine learning has the potential to be applied to more innovations. By automating some aspects of data analysis and prediction, engineers can devote more time to critical thinking, problem solving and creativity. This allows a shift in focus to pave the way for breakthroughs that were previously hampered by time-consuming tasks (Reich, 1997; Naranjo-Pérez et al., 2020).

This paper presents an estimation of the coefficient of permeability using various machine learning techniques, starting from the traditional predictive linear regression method and contrasting it with several more modern and developed machine learning techniques – the artificial neural network, the random forest and the gradient boosting. As a result, it verified the view that these techniques allow for more efficient estimation of the filter coefficient, with increasingly newer algorithms providing opportunities to reduce estimation errors.

MATERIAL AND METHODS

The analysis was carried out for sands with the grain sizes presented in Figure 1. The grain sizes were within the range shown in the figure. The study of the coefficient of permeability was performed using the constant head method. In addition, material properties were considered, such as volumetric density ranging from 0.99 g·cm^{-3} to 1.83 g·cm^{-3}, porosity 0.29 to 0.61 [–], index porosity 0.40 to 1.56 [–], grain size curvature index from 1.07 to 1.16 [–] and homogeneity index from 1.94 to 2.40 [–].

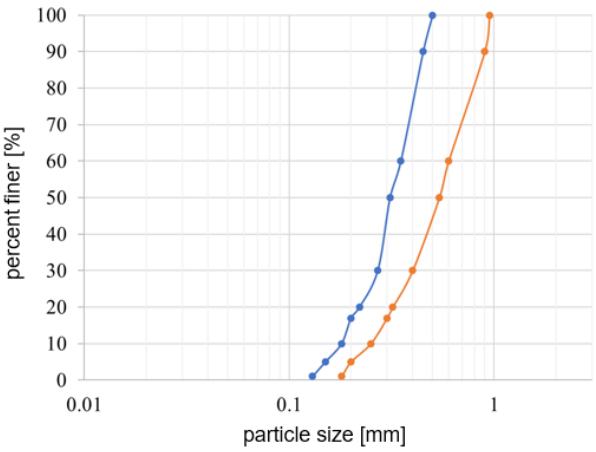

Fig. 1. Range grain size curves for the materials analysed

Source: own work.

One of the first and widely known predictive techniques is the least squares method – linear regression. It is still finding applications for the verification of simple feature relationships. Throughout the 1990s

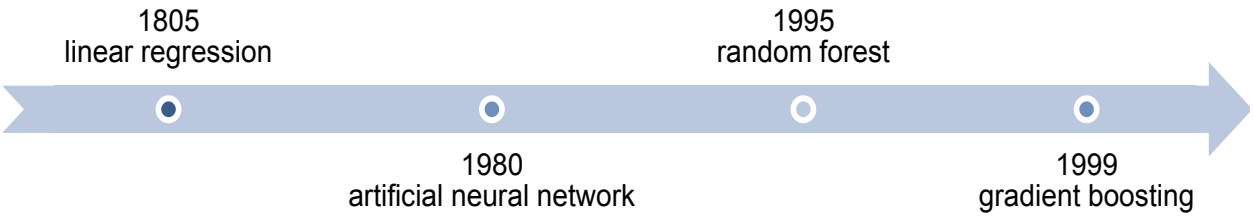

Fig. 2. Diagram of the formation of the machine learning techniques

Source: own work.

and the last two decades (Fig. 2), the development of the algorithm has significantly accelerated and can be said to be directly proportional to the development of computer techniques and capabilities. Several predictive techniques were used in the analysis, which are characterised further.

The linear regression

The linear regression is a statistical method employed to model the association between a dependent variable, also referred to as the target, and one or more independent variables, known as predictors or features. The principal objective of the linear regression is to determine the optimal linear equation that accurately characterises this relationship (Barbur, Montgomery & Peck, 1994; Weisberg, 2005; Seber & Lee, 2012). This equation takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \epsilon, \quad (1)$$

where:

y – dependent variable (target),
$x_1, x_2, \ldots, x_n$ – the independent variables (features),
$\beta_0, \beta_1, \ldots, \beta_n$ – the coefficients representing the impact of each feature on the target,
ϵ – the error term representing the difference between the predicted and actual values.

The primary aim of the linear regression is to estimate the coefficients ($\beta_0, \beta_1, \ldots, \beta_n$) which minimise the sum of squared differences between predicted and actual values of the dependent variable. This optimisation process is often accomplished through techniques like the least squares method. Linear regression operates under the assumption of a linear relationship between the features and the target. It finds applications across diverse domains, including prediction, correlation analysis and assessment of variable influence. Variants such as multiple linear regression (with multiple predictors) and polynomial regression (addressing certain nonlinear relationships) extend their adaptability (Hu et al., 2019; Maulud and Abdulazeez, 2020).

The artificial neural network

The artificial neural network (ANN) is a machine learning model inspired by the architecture of the human brain. It comprises interconnected nodes, referred to as neurons, arranged in layers: input, hidden (one or more) and output. Neurons are connected through weighted connections. The artificial neural networks undertake various tasks, encompassing regression, classification and pattern recognition. The process involves:

- Input layer: neurons represent data features.
- Hidden layers: neurons process inputs through mathematical operations, often involving weighted sums and activation functions. Hidden layers enable the network to capture intricate data relationships.
- Activation functions: neurons employ activation functions to introduce nonlinearity. Common functions include sigmoid, ReLU and tanh.
- Output layer: The final hidden layer connects to the output layer, generating predictions. The number of output neurons varies based on the task (e.g. regression, classification).
- Training: ANNs learn by adjusting connection weights to minimise a loss function. Back propagation computes gradients of loss concerning weights, guiding weight updates through optimisation techniques such as gradient descent.

The objective is to identify weights that minimise loss, involving iterative weight updates to reduce loss. Techniques like stochastic gradient descent are commonly employed.

The artificial neural networks are adept at modelling both linear and nonlinear relationships, providing flexibility for complex tasks. The deep neural networks (DNN), equipped with multiple hidden layers, excel in capturing intricate patterns. Designing and training ANNs necessitate careful consideration of architecture, hyperparameters and data to prevent issues like overfitting (Suzuki & Soleimanian Gharehchopogh, 2012; Lagaros, 2023).

The random forest

The random forest is the machine learning ensemble method used for both classification and regression tasks. It is based on the concept of decision trees and combines multiple individual decision trees to create a more robust and accurate predictive model. The random forest starts from creating multiple decision trees. To introduce diversity among these trees, each tree is trained on a randomly sampled subset of the training data, with replacement. This technique is known as bagging. In addition to sampling data, the random forest introduces randomness in feature selection. When creating each split in a decision tree, the algorithm considers only a random subset of the available features. This prevents individual trees from becoming overly specialised and reduces the risk of overfitting. The individual decision trees created using bagging and feature randomness combine to form the random forest ensemble. For regression tasks, the final prediction is often the average of predictions from all trees. For classification tasks, the ensemble's prediction can be determined by a majority vote among the individual trees. To predict a new data point, the input is passed through each individual tree in the random forest and the final prediction is aggregated according to the ensemble method – average for regression or majority vote for classification (Breiman, 2001; Cutler, Cutler & Stevens, 2012; Louppe, 2014).

The gradient boosting

The gradient boosting is an ensemble method utilised for regression and classification tasks. It assembles a potent model by amalgamating predictions from weak learners (often decision trees) in a sequential manner. The process commences with a rudimentary prediction (e.g. mean of target). It computes residuals, indicating the disparity between actual target values and initial predictions, then constructs trees sequentially to predict negative gradients of the loss function and introduces a learning rate parameter to scale tree predictions before their addition to the ensemble. A smaller learning rate fosters gradual and stable learning. New trees' predictions enhance the existing ensemble's performance. The boosting process continues until a predefined number of trees is attained or a specific performance threshold is reached.

The gradient boosting excels in capturing intricate relationships within (Friedman, 2002; Velthoen, Dombry, Cai & Engelke, 2021).

These algorithms are foundational tools within the realm of the machine learning, each offering distinct strengths and applications. A schematic of the estimation process using the machine learning algorithms is shown in Figure 3. Ten-fold cross-validation, a resampling technique for evaluating the machine learning models on limited data, was used to validate the model. The data (n = 261) was collected, cleaned and divided into 70% training samples and 30% test samples. Cross-validation helps estimate the model's predictive ability on unseen data. The k-fold parameter divides the data into groups, used to evaluate

Fig. 3. Diagram of the estimation process using the machine learning algorithms

Source: own work.

model performance. This method provides a less biased estimate of model capability (Browne, 2000).

The final model's reliability was confirmed through average skill scores and measures of variance. Error analysis was employed to evaluate individual model performance. The evaluation included the artificial neural network, the random forest and the gradient boosting algorithms; linear regression was used as a reference and control algorithm. The results were verified by error analysis and the following values were estimated for each model:

– coefficient of determination (R^2):

$$R^2 = \frac{\Sigma_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\Sigma_{i=1}^{N}(y_i - \overline{y})^2}, \tag{2}$$

– mean absolute error (*MAE*):

$$MAE = \frac{\Sigma_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n}, \tag{3}$$

– root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y})^2}. \tag{4}$$

RESULTS AND DISCUSSION

The estimation results for the learning and test samples are presented in Figure 4. Coefficient of determination, *MAE* and *RMSE* were calculated for each of the algorithms and each of the trials according to Eqs (2)–(4). Figure 4 presents plots of the accuracy of the fit of the estimation results to the data obtained from the trials. The lowest values of fit, R^2 were obtained for the linear regression algorithms and were equal to 0.833 for the learning sample and 0.800 for the test sample, the *MAE* was 0.003, the *RMSE* was 0.004; for the neural network, R^2 was 0.833 for the learning sample and 0.801 for test sample, the *MAE* was 0.003, and the *RMSE* was 0.004. The best results were obtained for the gradient boosting algorithm for both samples: R^2 was 0.995, the *MAE* was less than 0.001 and the *RMSE* was 0.001.

The reasons for differences in estimation accuracy should be sought in the characteristics of individual algorithms. Among other things, the type of task for which the algorithm is constructed may be of significance in the case of linear regression which is mainly used in regression tasks. In the case of the ANN algorithm, its application is comprehensive, because it can be used for regression and classification tasks. The same is true for the random forest and the gradient boosting. Another issue is handling: dealing with nonlinearity in the case of linear regression is limited to modelling linear relationships. The ANN perfectly captures both linear and nonlinear dependencies. The random forest and the gradient boosting effectively handle nonlinear relationships.

It is also important that the structure of the model linear regression is characterised by a simple structure involving a linear equation. ANN has a complex architecture with interconnected layers of nodes. Random forest forms a set of decision trees and the gradient boosting is a set of sequentially improved models. The architecture of the model affects its interpretability. The linear regression allows high interpretability due to the linear equation. The ANN is less interpretable due to its complex structure. Random forest offers insight into the meaning of the function. Gradient boosting is less interpretable compared to linear models. Performance and complexity are also not negligible, especially for complex estimation tasks. Linear regression is simple and computationally efficient; in the case of ANN the algorithm is designed for complex and computationally intensive issues. Random forest and gradient boosting are sustainable performances for solving complex problems, but these techniques can be computationally demanding. An important issue from the point of view of evaluating the correctness of the estimation is the control of overfitting. Linear regression is a technique prone to overfitting with complex relationships. The ANN is prone to overfitting, especially with small data sets. Random forest is immune to overfitting due to the nature and construction of the algorithm. Gradient boosting can cause overfitting but is less likely compared to single decision trees.

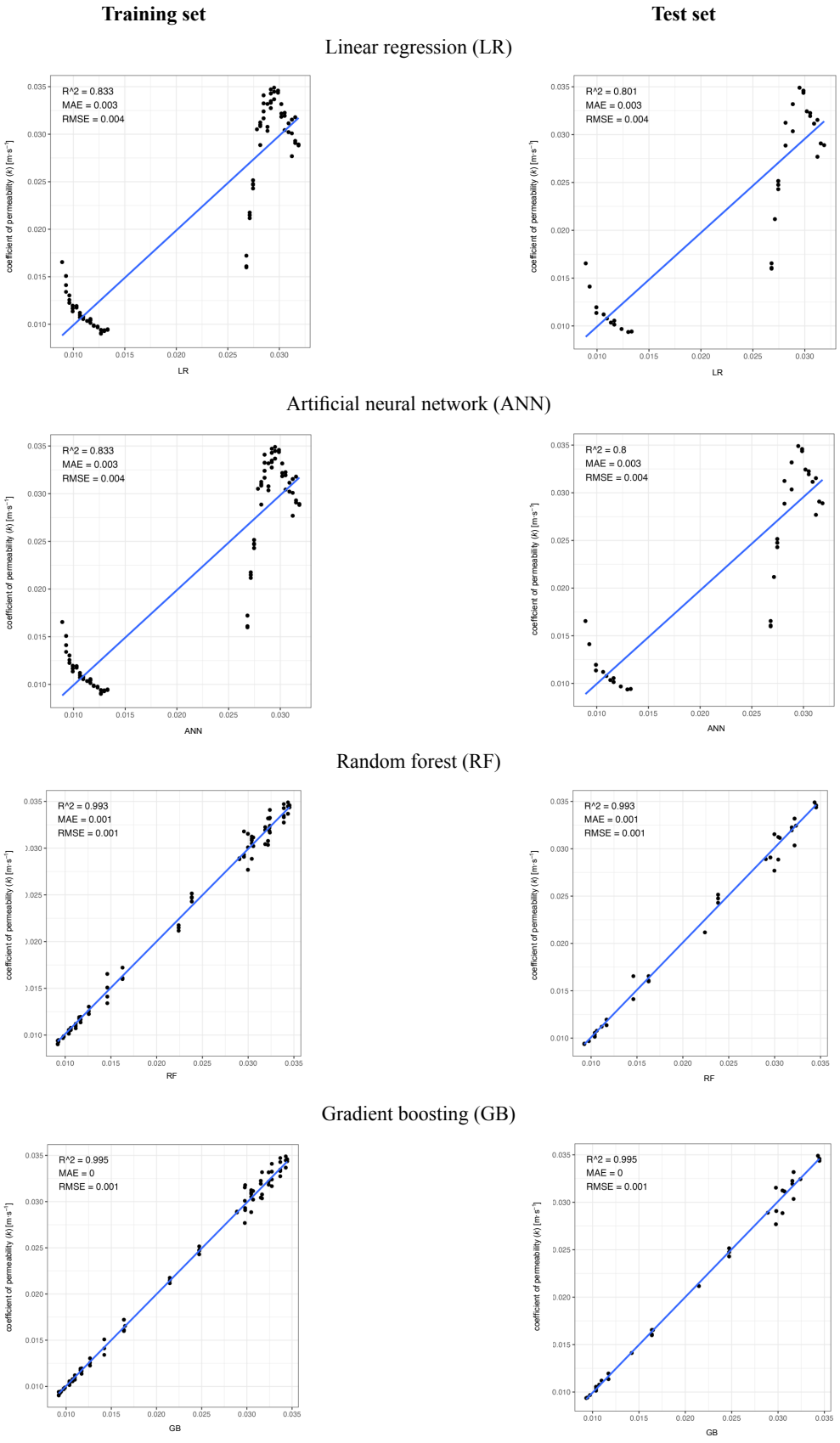

Fig. 4. Estimation results for the training and test set

Source: own work.

CONCLUSIONS

The machine learning techniques offer significant potential for predicting the permeability coefficient in civil engineering. By leveraging large datasets and complex algorithms, these methods provide more efficient and accurate predictions compared to available empirical formulas. The permeability coefficient prediction plays a crucial role in groundwater flow analysis, soil stabilisation and geotechnical engineering applications. However, further research and validation are necessary to ensure the reliability and applicability of the machine learning models in real-world civil engineering projects. The algorithms analysed in the article vary in capability, complexity and suitability for different types of data and tasks. The choice depends on factors such as data characteristics, interpretation requirements, performance expectations and available computing resources. In the case of the analysed soil – sands, the algorithm with the highest predictive efficiency turned out to be the gradent boosting whose matching of the prediction results with the data derived from laboratory tests amounted to 0.995. At the same time, it should be noted that to generalise, the data obtained should be analysed on a wider database and based on a larger number of materials.

REFERENCES

Barbur, V. A., Montgomery, D. C. & Peck, E. A. (1994). Introduction to Linear Regression Analysis. *The Statistician*, *43* (2), 339. https://doi.org/10.2307/2348362

Breiman, L. (2001). Random forests. *Machine Learning*, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324

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

Cutler, A., Cutler, D. R. & Stevens, J. R. (2012). Random forests. In *Ensemble Machine Learning: Methods and Applications* (pp. 157–175). Boston, MA: Springer. https://doi.org/10.1007/9781441993267_5

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38* (4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Hu, Y. H., Yu, S. C., Qi, X., Zheng, W. J., Wang, Q. Q. & Yao, H. Y. (2019). An overview of multiple linear regression model and its application. *Zhonghua Yu Fang Yi Xue Za Zhi* [*Chinese Journal of Preventive Medicine*], *53* (6), 653–656. https://doi.org/10.3760/CMA.J.ISSN.0253-9624.2019.06.021

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

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice* (PhD dissertation). University of Liège, Liège. https://doi.org/10.48550/arxiv.1407.7502

Maulud, D. & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1* (4), 140–147.

Melhem, H. G. & Nagaraja, S. (2007). Machine learning and its application to civil engineering systems. *Civil Engineering Systems*, *13* (4), 259–279. https://doi.org/10.1080/02630259608970203

Naranjo-Pérez, J., Infantes, M., Fernando Jiménez-Alonso, J. & Sáez, A. (2020). A collaborative machine learning-optimization algorithm to improve the finite element model updating of civil engineering structures. *Engineering Structures*, *225*, 111327. https://doi.org/10.1016/J.ENGSTRUCT.2020.111327

Kosinov, V., Trach, Y. & Trach, R. (2023). Analysis of the construction of nodes of a water pipeline network and modeling of planned overall dimensions of its working chambers. *Acta Scientiarum Polonorum. Architectura*, *21* (1), 71–80. https://doi.org/10.22630/ASPA.2022.21.1.8

Pardalos, P. M., Rassia, Th. S. & Tsokas, A. (Eds), (2022). *Artificial Intelligence, Machine Learning and Optimization Tools for Smart Cities* (Springer Optimization and Its Applications. Vol. 186). Retrieved from: https://link.springer.com/10.1007/978-3-030-84459-2 [accessed: 07.09.2023].

Reich, Y. (1997). Machine Learning Techniques for Civil Engineering Problems. *Computer-Aided Civil and Infrastructure Engineering*, *12* (4), 295–310. https://doi.org/10.1111/0885-9507.00065

Seber, G. A. F. & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). Retrieved from: https://books.google.com/books/about/Linear_Regression_Analysis.html?hl=pl&id=X2Y6OkXl8ysC [accessed: 07.09.2023].

Suzuki, K. & Soleimanian Gharehchopogh, F. (2012). Artificial Neural Networks – Methodological Advances and Biomedical Applications. In *Artificial Neural Networks – Methodological Advances and Biomedical Applications*. https://doi.org/10.5772/644

Szymanek, S. (2022). Construction Production Trends and Industry Optimism in EU Countries After The COVID-19 Pandemic. *Acta Scientiarum Polonorum.*

Architectura, *21* (4), 69–74. https://doi.org/10.22630/ASPA.2022.21.4.32

Velthoen, J., Dombry, C., Cai, J-J. & Engelke, S. (2021). *Gradient boosting for extreme quantile regression*. https://doi.org/10.48550/arxiv.2103.00808

Weisberg, S. (2005). *Applied linear regression* (wyd. 3). Hoboken, NJ: John Wiley & Sons. Retrieved from: https://books.google.com/books/about/Applied_Linear_Regression.html?hl=pl&id=xd0tNdFOOjcC [accessed: 07.09.2023].

UCZENIE MASZYNOWE W INŻYNIERII LĄDOWEJ NA PRZYKŁADZIE PRZEWIDYWANIA WSPÓŁCZYNNIKA PRZEPUSZCZALNOŚCI

STRESZCZENIE

W niniejszym opracowaniu zbadano zastosowanie technik uczenia maszynowego w inżynierii lądowej, koncentrując się na przewidywaniu współczynnika przepuszczalności. Współczynnik przepuszczalności jest istotnym parametrem w różnych projektach inżynierii lądowej, takich jak: analiza przepływu wód gruntowych, stabilizacja gruntu i inżynieria geotechniczna. Tradycyjne metody szacowania przepuszczalności są czasochłonne i często opierają się na testach laboratoryjnych. Uczenie maszynowe oferuje obiecujące podejście do jego przewidywania w sposób bardziej wydajny i dokładny. W niniejszym artykule przeanalizowano kilka technik uczenia maszynowego, weryfikując możliwość ich zastosowania do przewidywania współczynnika przepuszczalności dla piasków. W artykule przeanalizowano skuteczność predykcyjną artificial neural network (ANN), random forest (RF), gradient boosting (GB) i regresji liniowej (LR). Najdokładniejszym algorytmem w tym wypadku okazał się GB, dla którego współczynnik determinacji wyniósł 0,995, średni błąd bezwzględny był na poziomie poniżej 0,001, a błąd średniokwadratowy wyniósł 0,001.

Słowa kluczowe: uczenie maszynowe, współczynnik filtracji, predykcja