PERMEABILITY COEFFICIENT. ESTIMATION TECHNIQUES FROM LINEAR REGRESSION TO MACHINE LEARNING ALGORITHMS - ON THE CASE OF BLAST FURNACE SLAG

Justyna Dzięcioł [1*] and Wojciech Sas [1]
[1] Institute of Civil Engineering, Warsaw University of Life Sciences – WULS, Warsaw, Poland
[*] E-mail: justyna_dzieciol@sggw.edu.pl, wojciech_sas@sggw.edu.pl

ABSTRACT
The exploration of alternatives which could replace natural aggregates in the construction sector has led to a change in thinking also about waste, which is increasingly being used in Civil and Road Engineering. These include Blast Furnace Slag which is a by-product of the steel production process. It finds a "second life" as an aggregate or concrete additive, for example. The application of anthropogenic aggregates as substitutes for natural aggregates offers the possibility of reducing the number of landfills. Meanwhile, it is important to effectively manage the material, which involves reliable estimation of its parameters. The main factor affecting the correctness of parameter estimation for anthropogenic aggregates is the fact that they can have different properties and chemical composition, depending on where they are manufactured. The coefficient of permeability for aggregates is an important parameter for the application of these materials in a variety of fields, such as construction, road infrastructure, and land reclamation. It determines the ability of aggregates to permeate water, which is crucial for pavement design. Therefore, correct estimation of the coefficient of permeability can significantly influence the stability and durability of structures. The objective of the study was to analyze various methods of estimating the coefficient of permeability beginning with the simple and well-known linear regression and compare it with other estimation methods and approaches, including with Machine Learning Algorithms. The results obtained were compared with each other, and based on them an attempt was carried out to interpret the physical characteristics of the aggregate, which had a significant impact on the estimation of the model. The analysis allowed to formulate conclusions and set directions for the selection of estimation techniques for the prediction of the permeability coefficient and other geotechnical parameters.

1. Introduction

The modern approach to the investment process is a key area of interest, requiring a balance between economic goals and environmental concerns. Sustainable design, encompassing the stages of construction, operation, and dismantling of a structure, has become an imperative for designers, engineers, and architects. The priority lies in minimizing resource consumption and selecting materials with a low environmental impact [1]. However, the dynamic development of construction infrastructure generates a significant demand for non-renewable aggregates, which may lead to limitations in access to deposits in the future [2].

In the context of ecological efficiency, the construction and operation of buildings not only consume large amounts of natural resources but also contribute significantly to greenhouse gas emissions. Forecasts indicate increasing challenges related to access to natural aggregates, further emphasizing the need to explore alternative solutions [2]. In the context of the European Union, which is a highly industrialized economy, there is an urgent need for the development of recycling technologies, especially considering that only 38% of all generated waste is currently subject to recycling [3].

One potential solution in the waste management sector is slag, a byproduct of industrial technological processes. In 2019, slag alone accounted for 14.3 million tons of industrial waste (Central Statistical Office, 2020). In the context of sustainable design, attention should be drawn to blast furnace slag (BFS), which undergoes a cooling process leading to the crystallization of minerals. Due to growing challenges in accessing natural aggregates, slag can serve as a valuable resource in the construction process [5, 6].

Machine learning techniques have their origins in the method of least squares, the first predictive method allowing for simplified estimation of linear parameters.

Linear Regression, one of the simplest methods for approaching regression problems, involves predicting an unknown variable using existing results. In the described method, the presence of residual values (residuals) is taken into account [7].

Artificial Neural Networks (ANN), also known simply as neural networks, represent a popular machine learning technique for data analysis through layers of decision-making. Neural networks divide data into layers and process hidden layers to obtain the final output. As additional hidden layers are added to the network, the model's ability to analyze complex patterns improves. Therefore, models with a large number of layers are often referred to as deep learning to highlight their deeper and enhanced processing capabilities [8].

Random Forest is an extension of decision tree techniques. Closely related to bagging (a method for optimizing machine learning algorithms), both techniques involve the application of multiple decision trees and utilize bootstrap sampling (a technique for aggregating results from multiple model estimations) for data randomization. Random forests artificially limit the selection of variables by restricting the number of variables considered for each split. In the case of bagging, decision trees often look similar because they use the same variable at the beginning of their decision structure to reduce entropy [9–11].

Gradient Boosting selects variables that improve prediction accuracy with each new tree. Decision trees grow sequentially, as each tree is created using information from the previous tree, rather than independently. Errors made in training data are recorded and then applied to the next round of training data. In each iteration, weights are added to the training data based on the results of the previous iteration [12, 13].

The article focuses on the application of machine learning algorithms to predict one of the most significant geotechnical parameters – the filtration coefficient. This parameter finds application in structures for water retention and in determining filtration through the soil of embankment constructions. Knowledge of the filtration coefficient is useful for selecting various filters, including reverse filters preventing adverse phenomena during water flow through a porous medium. The filtration coefficient also plays a significant role in road engineering, particularly in designing embankment layers for road constructions. The article also employs the SHAP interpretative technique to interpret physical material parameters that have a significant impact on determining the filtration coefficient.

2. Materials and Methodology

2.1 Materials

The Blast Furnace Slag used in the study came from a steel melting plant. The constant head method was used to test permeability characteristics for Blast Furnace Slag. The method is characterized by simplicity and unchanging test conditions, and the constant head method alone is one of the most reliable techniques for measuring permeability in non-cohesive soil [14, 15]. The coefficient of permeability study used aggregate of Blast Furnace Slag tested samples were from several parties of the material. Basic data on the grain size ranges of the tested samples are presented in Figure 1. and data on physical parameters are included in Table 1.

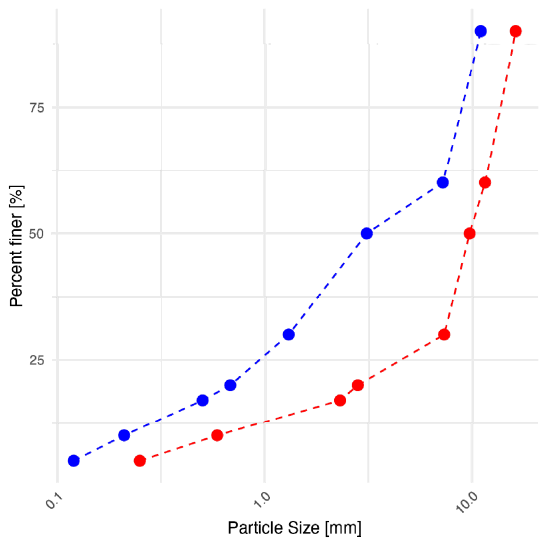

Figure 1. *Grain size curves of the tested material.*

The samples were compacted with an Proctor normal energy [16] of 0.59 [J/cm^3] and 2.65 [J/cm^3].

Table 1. *Physical parameters of Blast Furnace Slag.*

	Volumetric density	Porosity	Index porosity	Homogeneity index - Cu	Grain size curvature index - Cc	Gradient
Min.	1.02	0.54	1.18	19.49	1.12	0.08
Max.	1.08	0.57	1.31	34.29	7.85	1.02

2.2 Methodology

The process and methodology of the analysis are outlined in Figure 2. Prior to commencing the analysis, data collection and cleaning were undertaken. During the learning and testing phases, the data was partitioned into a training set (70%) and a test set (30%). Model validation was executed using the 10-fold Cross Validation method.

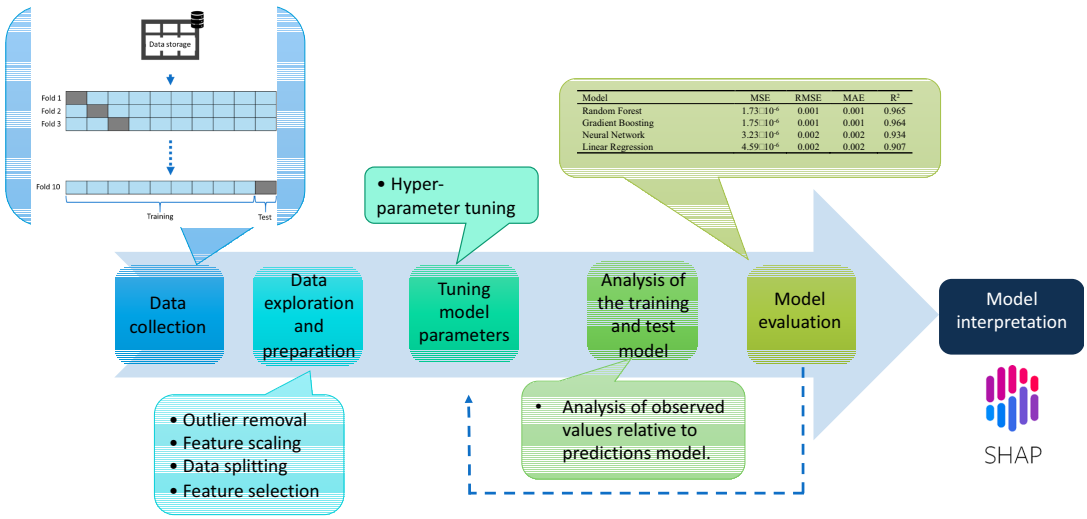

Figure 2. *Analysis scheme.*

Cross-validation, a resampling procedure, serves to evaluate machine learning models on a limited data sample. The procedure involves a parameter, k, denoting the number of groups in which the data sample is to be divided. When a specific value for k is chosen, it can be employed for the model; for instance, k=10 results in 10-fold cross-validation. This technique is widely utilized in Machine Learning to estimate a model's predictive ability. Essentially, it uses a restricted sample to gauge how well the model is expected to perform overall when making predictions on data not employed in the model's training. This method is favored for its simplicity and its tendency to provide a less biased or optimistic estimate of the model's ability compared to other methods. Incorporating a measure of variance in skill scores, such as standard deviation or standard error, is recommended for a comprehensive evaluation. The outcome of k-fold cross-validation is summarized by the average of the model's skill scores. This practice enhances the reliability of the model obtained during the learning and testing phases.

Including a measure of variance in skill scores further refines the assessment, contributing to a more nuanced understanding of the model's performance. The results were verified with the use of error analysis, for individual models were estimated:

- Mean Square Error (MSE)

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

- Root Mean Square Error (RMSE):

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

- Mean Absolute Error (MAE):

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

- Coefficient of determination (R^2):

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

Once the models for each Machine Learning algorithm were determined, an interpretive analysis was performed using the SHAP technique.

3. Results

The results chapter is divided into a part related to the discussion of the results of estimating the permeability coefficient using Machine Learning algorithms, the second part discusses the results of interpreting the various models and physical parameters of the material having the most significant impact on the estimation of the permeability coefficient.

3.1 Estimation results

The results of estimating the permeability coefficient using the four estimation techniques of linear regression, Random Forest, Gradient Boosting and Neural Network are presented below in Tables 2 and 3. The best results for the coefficient of determination, R^2, were obtained for the Random Forest algorithm. For this algorithm, R^2 = 0.965 was obtained for the learning set, and for the test set R^2 = 0.983. Similar results were obtained for the Gradient Boosting algorithm - R^2 = 0.964 was obtained for the learning set, and for the test set R^2 = 0.982. The lowest results of the learning set were obtained for Linear Regression, R^2 was 0.907, and for the test set for the Neural Network algorithm 0.790.

Table 2. *Evaluation on training set.*

Model	MSE	RMSE	MAE	R^2
Random Forest	1.73×10^{-6}	0.001	0.001	0.965
Gradient Boosting	1.75×10^{-6}	0.001	0.001	0.964
Neural Network	3.23×10^{-6}	0.002	0.002	0.934
Linear Regression	4.59×10^{-6}	0.002	0.002	0.907

Table 3. *Evaluation on test set.*

Model	MSE	RMSE	MAE	R^2
Random Forest	8.30×10^{-7}	0.001	0.001	0.983
Gradient Boosting	8.79×10^{-7}	0.001	0.001	0.982
Linear Regression	4.55×10^{-6}	0.002	0.002	0.907
Neural Network	1.02×10^{-5}	0.003	0.002	0.790

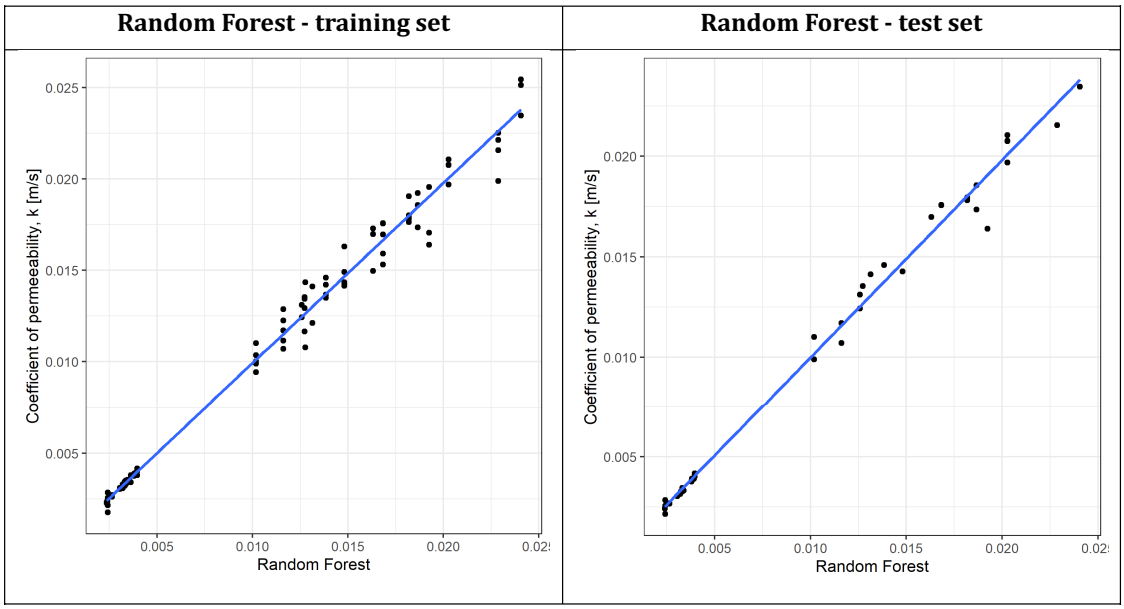

Figure 3. *Results of the permeability coefficient and predictive performance of Random Forest.*

The results of the permeability coefficient study and their comparison with the estimation results obtained for the algorithm with the highest predictive performance Random Forest are presented in Figure 3.

3.2 Interpretive analysis

The application of the SHAP interpretation technique allowed the determination of the physical properties of the material having the greatest influence on the formation of individual models based on the algorithms. The formation of the model based on the Random Forest algorithm was most influenced by Compaction energy, gradient and particle size, d50. Other model interpretation results are presented in Figure 4.

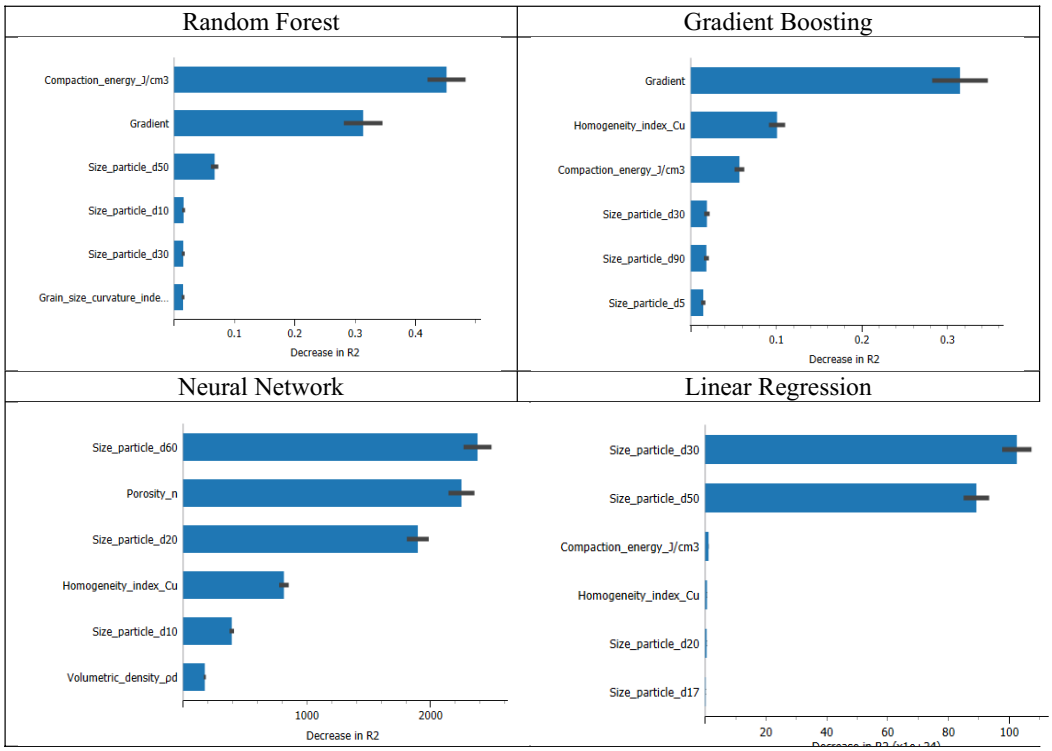

Figure 4. *The models interpretation results.*

4. Conclusion

The paper analyzes applications of Machine Learning algorithms to predict the permeability coefficient, a critical geotechnical parameter. In particular, Random Forest emerges as the most effective algorithm, achieving the highest coefficient of determination (R^2) in both the learning set (0.965) and the test set (0.983). Gradient Boosting also performed well in predicting this parameter, with R^2 values of 0.964 and 0.982 for the learning and test sets, respectively. Linear regression and Neural Network algorithms show lower R^2 values, indicating relatively less accurate predictions. The SHAP interpretation technique has been used to uncover influential physical material properties affecting individual models. The analysis reveals that factors such as compaction energy, gradient and particle size (d_{50}) significantly affect Random Forest model formation. In conclusion, the paper highlights the effectiveness of Machine Learning, particularly Random Forest, in predicting the filtration rate. SHAP interpretation provides valuable insight into the material properties that shape predictive models, enhancing the understanding of geotechnical parameters in the context of sustainable design and construction practices.

References

[1] Chen Y, Okudan GE, Riley DR. Sustainable performance criteria for construction method selection in concrete buildings. *Autom Constr* 2010; 19: 235–244.
[2] Kumar S, Skariah Thomas B, Gupta V, et al. Sandstone wastes as aggregate and its usefulness in cement concrete – A comprehensive review. *Renewable and Sustainable Energy Reviews* 2018; 81: 1147–1153.
[3] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. A new Circular Economy Action Plan. For a cleaner and more competitive Europe. COM(2020), https://www.un.org/sustainabledevelopment/sustainable-consumption-production/ (2020, accessed 23 October 2022).
[4] Główny Urząd Statystyczny (Central Statistical Office). *Statistical analyses, Environment 2020*, https://stat.gov.pl/obszary-tematyczne/srodowisko-energia/srodowisko/ochrona-srodowiska-2020,1,21.html (2020).
[5] Sas W, Dzięcioł J, Radzevičius A, et al. Geotechnical and environmental assessment of Blast Furnace Slag for engineering applications. *Materials* 2021; 14: 6029.
[6] Dzięcioł J, Radziemska M. Blast Furnace Slag, Post-Industrial Waste or Valuable Building Materials with Remediation Potential? *Minerals* 2022; 12: 478.
[7] Barbur VA, Montgomery DC, Peck EA. Introduction to Linear Regression Analysis. *The Statistician* 1994; 43: 339.
[8] Shanthamallu US, Spanias A. *Machine and Deep Learning Algorithms and Applications*. 2022. Epub ahead of print 2022. DOI: 10.1007/978-3-031-03758-0.
[9] Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
[10] Genuer R, Poggi J-M, Tuleau C. Random Forests: some methodological insights. Epub ahead of print 21 November 2008. DOI: 10.48550/arxiv.0811.3619.
[11] Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat* 2019; 47: 1179–1203.
[12] Manoharan A, Begam KM, Aparow VR, et al. Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review. *J Energy Storage* 2022; 55: 105384.
[13] Velthoen J, Dombry C, Cai J-J, et al. Gradient boosting for extreme quantile regression. Epub ahead of print 1 March 2021. DOI: 10.48550/arxiv.2103.00808.
[14] Sas W, Dzięcioł J, Głuchowski A. Estimation of recycled concrete aggregate's water permeability coefficient as earth construction material with the application of an analytical method. *Materials* 2019; 12: 2920.

[15] Sas W, Dzięcioł J. Determination of the filtration rate for anthropogenic soil from the recycled concrete aggregate by analytical methods. *Scientific Review Engineering and Environmental Sciences* 2018; 27: 236–248.
[16] PN-EN 13286-2:2007. *Unbound and hydraulic binder mixtures: Part 2: Methods for determining density in relation to water content. Proctor compaction.* 2007.