minerals

MDPI

Article

Prediction of the Adsorption Behaviors of Radionuclides onto Bentonites Using a Machine Learning Method

Do-Hyeon Kim and Jun-Yeop Lee *

School of Mechanical Engineering, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea
* Correspondence: jylee@pusan.ac.kr

Abstract: This study builds a model to predict distribution coefficients (K_d) using the random forest (RF) method and a machine learning model based on the Japan Atomic Energy Agency Sorption Database (JAEA-SDB). A database of ten input variables, including the distribution coefficient, pH, initial radionuclide concentrations, solid–liquid ratio, ionic strength, oxidation number, cation exchange capacity, surface area, electronegativity, and ionic radius, was constructed and used for the RF model calculation. The calculation parameters employed in this work included two different hyperparameters, the number of decision trees and the maximum number of variables to divide each node, together with the random seeds inside the RF model. The coefficients of determination were derived with various combinations of hyperparameters and random seeds, and were employed to assess the RF model calculation result. Based on the results of the RF model, the distribution coefficients of 22 target nuclides (Am, Ac, Co, Cm, Cd, Cs, Cu, Na, Np, Ni, Nb, U, Sr, Sn, Pb, Pa, Pu, Po, I, Tc, Th, and Zr) were predicted successfully. Among the various input variables, pH was found to make the highest contribution to determining the distribution coefficient. The novelty of this study lies in the first application of the machine learning method for predicting the K_d value of bentonites, using JAEA-SDB. This study has established a model for reliably predicting the distribution coefficient for various radionuclides that is intended for use in evaluating the K_d value in arbitrary aqueous conditions.

Keywords: adsorption; bentonite; distribution coefficient; machine learning; random forest

Citation: Kim, D.-H.; Lee, J.-Y. Prediction of the Adsorption Behaviors of Radionuclides onto Bentonites Using a Machine Learning Method. *Minerals* **2022**, *12*, 1207. https://doi.org/10.3390/min12101207

Academic Editors: Honty Miroslav, Stephan Kaufhold, Ana María Fernández and Patrik Sellin

Received: 20 June 2022
Accepted: 21 September 2022
Published: 25 September 2022

1. Introduction

Due to the continuous use of nuclear power, the generation and storage of spent nuclear fuel are expected to increase progressively. Since these radiotoxic high-level wastes (HLW) contain inherent radiological hazards to the environment and to humans, it is urgently necessary to establish a safe and reliable disposal plan for the HLW.

Recently, the KärnBränsleSäkerhet-3 (KBS-3) deep geological disposal concept [1], developed by the Svensk Karnbranslehantering AB (SKB) in Sweden, was adopted and applied to the design and safety assessment of the ONKALO final repository for spent nuclear fuel in Finland, which is currently under construction. In the KBS-3 concept, bentonite-based buffer material is placed outside the copper-cast iron canister and functions as an engineered barrier to retard the migration of radionuclides released from the canister into the geosphere. In this respect, the primary retention process of radionuclide migration by the bentonite can be described by the adsorption mechanism, which can be quantified simply with a distribution coefficient (K_d). The distribution coefficient is a conditional constant, which means that it is highly dependent on the given geochemical variables, such as the pH, solid–liquid ratio (L/S ratio), temperature, species distribution, etc. Therefore, for a reliable assessment of the long-term safety of HLW disposal, it is essential to examine and derive the exact adsorption distribution coefficients by which the site-specific geochemical conditions are appropriately reflected. In this framework, the Japan Atomic

Energy Agency (JAEA) has developed a sorption database (JAEA-SDB) summarizing and displaying the extensive distribution coefficients of the major radionuclides that constitute the radioactive waste on important mineralogical matters, which were selected through a meticulous reliability evaluation of the reported values in many previous studies. The JAEA-SDB contains about 70,000 pieces of adsorption data, including information on the adsorbent material, target radionuclide, concentration, oxidation number, pH, distribution coefficient, etc. [2].

However, due to the relatively high uncertainty and conditional dependency of the distribution coefficient, it is challenging to predict the adsorption behaviors of radionuclides unless they have been investigated experimentally. While several thermodynamic model-based approaches (e.g., the surface complexation model [3]) have been developed to overcome such difficulties, it is still the case that limited fundamental chemical thermodynamic data are available to quantitatively assess the distribution coefficient of various radionuclides, leaving a significant uncertainty in terms of the prediction of adsorption behaviors.

Machine learning is a field of computing algorithms that has rapidly developed in recent years and is designed to imitate human intelligence by learning reams of data for establishing prediction, classification, and regression models [4]. Based on the development of computer CPUs, improvements in memory speed, and an increase in computing power, machine learning has been evaluated as an effective methodology for predicting specific values. In this regard, various techniques have been developed and used accordingly [5–8]. Machine learning can be divided into two approaches: supervised learning and unsupervised learning. Typically, supervised learning is used in learning data and for creating a model that can predict the desired target value. Machine learning approaches, such as the artificial neural network (ANN), random forest (RF), and support vector machine (SVM), are regarded as the established methodologies for general supervised learning [9,10]. Among them, the RF is an ensemble machine learning technique based on decision trees, which uses a voting method for classification and prediction with diverse decision trees. In the RF approach, various decision trees are created using random variables for prediction; the results are obtained through voting and averaging by collecting the results from the trees [11]. Furthermore, the K-fold cross-validation method can be additionally employed to avoid possible overfitting of the training data [12]. However, because of the data leakage and bias problems possibly induced by the K-fold cross-validation calculation, the nested K-fold cross-validation can be alternatively employed to provide better robustness and model stability by using training, validation, and test sets split from the original data set [13,14].

This study aims to reliably predict the distribution coefficients of primary radionuclides onto several bentonite materials, based on the RF machine learning model, with the JAEA-SDB as a data source. Furthermore, it is intended to quantify the relative influence and importance of various input variables on the distribution coefficient, by overcoming the limitations of human intuition and current chemical thermodynamic model-based approaches through machine learning models [15], allowing more reliable follow-up supplementary experiments.

2. Materials and Methods

2.1. Data Collection and Pre-Processing

The distribution coefficient data used for machine learning were taken from the JAEA-SDB [2]. Three major bentonites, comprising MX-80 [16], Kunigel V1 [16], and SWy-2 [17], which have been widely considered as buffer materials in the deep geological repository, were selected as the representative adsorbent materials in this study. The establishment of the machine learning database was achieved by using the 777 experimental results related to the 22 kinds of radionuclides that are provided by the JAEA-SDB (i.e., Am, Ac, Co, Cm, Cd, Cs, Cu, Na, Np, Ni, Nb, U, Sr, Sn, Pb, Pa, Pu, Po, I, Tc, Th, and Zr). The data used to build the RF model from JAEA-SDB included six geochemical variables: the solid-liquid ratio (mL/g, expressed as LS in the future), ionic strength (mol/L, referred to

as Ionic_S), the oxidation number of radionuclides (referred to as Redox), acidity (referred to as pH), initial radionuclide concentration (mol/L, referred to as C_0), and the distribution coefficient (m^3/kg, referred to as K_d). In addition, cation exchange capacity (meq/100 g, referred to as CEC) and surface area (m^2/g, referred to as SA), which are the characteristic properties of bentonite, along with electronegativity (referred to as EN) and ionic radius (Å, referred to as IR), which are unique features of radionuclides, were adopted in the modeling as well. The target value of machine learning was set to K_d (m^3/kg). For the simplification of the RF model and collective comparison among the assorted radionuclides, the ionic radii of radionuclides with a coordination number of 6 were established from the literature [18]. In addition, electronegativity values were also referenced from the previous study [19]. The CEC and SA data were entered separately, according to bentonite type (i.e., MX-80 [16], Kunigel V1 [16], and SWy-2 [17]). Furthermore, when the background solution was set to "seawater" in JAEA-SDB, the ionic strength was estimated and entered by assuming a standard seawater condition according to Millero et al. [20]. Likewise, the ionic strength for each scenario was calculated and entered collectively, following the reports of Torstenfelt et al. [21,22] for the background solution, labeled as "synthetic groundwater", Mateus et al. [23] for "tap water", and Kitamura et al. [24] for "synthetic porewater". The detailed input data constructed in the present work are provided in the Supplementary Materials.

For the machine learning database, constructed as above, the Pearson correlation coefficient (PCC) method was used to confirm the linear dependence between each variable [25]. The PCC is a method that can quantify the correlation between two arbitrary variables, which can be expressed as below:

$$r_{xy} = \frac{\left(\frac{\sum_{i=1}^{n}(x_i-\hat{x})\sum_{i=1}^{n}(y_i-\hat{y})}{n-1}\right)}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\hat{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-1}}} \quad (1)$$

where $\hat{x}$ and $\hat{y}$ are the average values of x and y, respectively. If the r_{xy} value is close to ± 1, the variables x and y have an almost direct linear relationship. In this case, eliminating one of the variables may be more advantageous for simplifying the model and reducing calculation time. This study evaluated the appropriateness of variable selection by calculating the correlation between variables using the PCC method, before performing the machine learning calculation.

2.2. Machine Learning Model

Random forest (RF) is a supervised ensemble machine learning approach, based on multiple decision trees and bagging (i.e., bootstrap aggregation) [11]. In general, the procedure of RF calculation can be expressed as follows:

1. The retained input data are randomly divided into a training set and a test set;
2. The RF model is created, followed by setting the hyperparameters used to control the learning process [26];
3. Multiple decision trees are created while training the model with the training set through each node;
4. The final result was estimated by averaging the results from all trees generated.

By adjusting the random state variable, i.e., a random seed in the RF calculation, randomness can be given in the classification of training/testing sets and the calculation using hyperparameters. The calculation parameters used for the RF calculation in this study include two different hyperparameters: the number of decision trees (N_{tree}) and the maximum number of variables to be used to divide each node ($N_{feature}$), together with the random state, random seed in the RF model [27]. In particular, the random seed numbers used to distinguish the training and test sets are referred to as "Random state_T", and the random seed for internal calculation of the RF model was referred to as "Random state_M". In the present work, the calculation parameters are set separately to calculate and assess

the RF model result. Further detailed information about the algorithm of the RF model is provided elsewhere [11].

Since the RF calculation makes decision trees by directly comparing the training set that is used for training with the test set, the overfitting of data can presumably be expected. Therefore, cross-validation using the K-fold cross-validation method can be utilized to check the calculation result. In this study, to secure more robustness and model stability, the nested K-fold cross-validation method was employed to additionally validate the RF result. The nested K-fold cross-validation consists of two loops, i.e., the inner fold and the outer fold. For the outer K-fold calculation, the given data set is divided into the K total number of equal parts. Then, one of the K parts is designated for use in evaluating the performance of the model, while the others are used as a training set. Successively, the K-1 equal parts are divided into J total equal parts. Similarly, one of the J parts is chosen for use as a validation set, while the other J-1 parts are used as a sub-training set. For the inner J-fold calculation, cross-validation is carried out with the sub-training set and the validation set to identify and select the calculation parameter combination (in this case, N_{tree}, $N_{feature}$, and Random state_M) providing the best performance. Note that the Random state_T, the random seed for dividing the training, testing, and validation set is fixed to be a constant value during the entire inner and outer loop calculation. In the present work, the performance of the inner loop model is evaluated with the averaged R^2 value obtained in the J-fold calculation. Sequentially, the calculation parameter combination selected through the inner fold calculation is employed to train the training set in the outer loop. Furthermore, the performance evaluation of the outer loop is conducted with the test set, and the entire process is repeated a total of K times. The model performance is assessed by averaging the number of K results obtained through the K-fold calculation at a given Random state_T value. In this study, the 5-5 nested cross-validation with the five inner loops and the five outer loops was employed. Additionally, the whole calculation process can be repeated with various Random state_T values to check the stability of the machine learning model. Through the above process, the best values of N_{tree} and $N_{feature}$ for the different Random state_M and Random state_T values can be identified. Finally, the representative values of N_{tree} and $N_{feature}$ are determined and fixed by (i) selecting the most frequent $N_{feature}$ values and (ii) averaging the best N_{tree} values, respectively. The final model test can be performed by averaging the R^2 results obtained with the representative N_{tree} and $N_{feature}$ values, together with various Random state_M and Random state_T values.

Additionally, the RF model quantitatively evaluates the relative importance of the input variables on the distribution coefficient based on the mean decrease in impurity (MDI) [15] approach. This method defines the average impurity reduction as an essential factor when dividing nodes, through which the relative importance of each variable used in the RF calculation can be compared.

Finally, the results of the RF model were assessed using the coefficient of determination (R^2) and the root mean square error (RMSE), calculated according to Equations (2) and (3) [28,29]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(Y_{i,\,exp} - Y_{i,\,pred} \right)^2}{\sum_{i=1}^{N} \left(Y_{i,\,exp} - \hat{Y}_{ave,\,exp} \right)^2} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \left(Y_{i,\,exp} - Y_{i,\,pred} \right)^2} \tag{3}$$

where $Y_{i,exp}$ is the actual experimental value and $Y_{i,pred}$ represents the value predicted by the RF model. As the R^2 value converges to 1 or RMSE to 0, the predicted data more closely agrees with the experimental data. The computational codes for the RF model calculation and the nested K-fold cross-validation employed in this study were taken from the scikit-learn software package (version 0.24.2) [30–32] and are provided in the Supporting Information.

3. Results

3.1. Data Processing and PCC Analysis

Among the input data used in this study, the relative distributions of K_d and C_0 were analyzed by plotting the boxplot diagram. As shown in Figure 1a, the K_d values represent a wide but relatively uneven distribution, ranging from 8.0×10^{-5} m^3/kg to 2.1×10^3 m^3/kg. As for the overall distribution of K_d values, small K_d values accounted for the most of data, while the K_d values larger than 500 m^3/kg occupy a much smaller fraction.

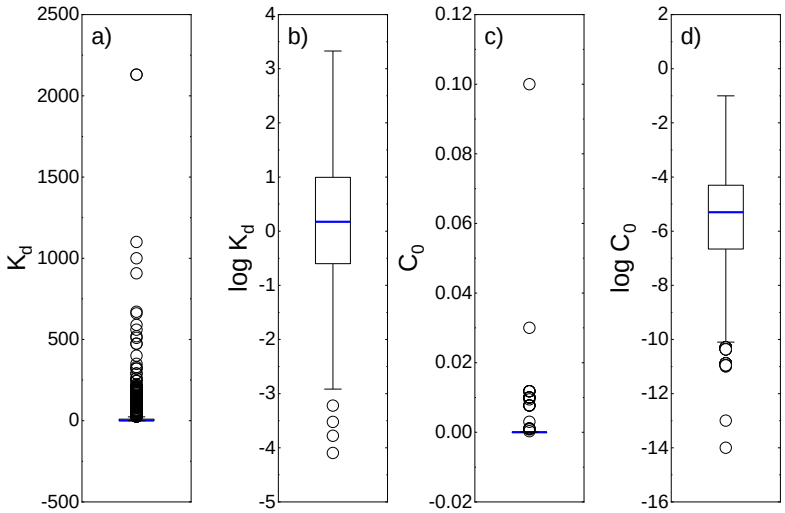

Figure 1. Box plots of the input variables, such as (**a**) K_d, (**b**) log K_d, (**c**) C_0, and (**d**) log C_0 employed in the RF model calculation. The blue bar indicates the median.

Figure 1b indicates that if the logarithmic scale was applied to the K_d value, the overall range of the values was reduced to an approximate range from −4.1 to +3.3, with a relatively uniform distribution. At the same time, the initial radionuclide concentration, C_0, showed a relatively consistent trend compared to the distribution coefficient, K_d (see Figure 1c,d).

Besides this, the solid-liquid ratio, LS, also presented an even distribution in the logarithmic scale compared to the linear scale, as with the K_d (data not shown). Therefore, this study used log K_d, log C_0, and log LS as the input training data for RF modeling, to prevent the overall calculation results from being biased by a small number of data values with large numerical values.

Figure 2 shows the data obtained by calculating the relationship between each input variable pair using the Pearson correlation coefficient (PCC) method. According to the results, there was no case showing a correlation that was close to ±1 for all the possible combinations. Therefore, all the nine input variables described in the Materials and Methods section, along with the log K_d, which represents the predicted target value of the RF model, were employed for the machine learning calculations.

In the Pearson correlation matrix result, the variables representing a remarkable correlation with log K_d were determined to be pH, log C_0, EN, and log LS. Thus, it was predicted that these four variables would show relatively high importance in predicting the K_d values through the machine learning algorithm.

3.2. RF and 5-5 Nested Cross-Validation

In this study, the calculation parameters controlled in the RF model were two different hyperparameters (N_{tree} and $N_{feature}$) and two different random seeds (Random state_T and Random state_M). The ranges of the calculation parameters were as follows:

- N_{tree} = 5–1000 (set in multiples of five intervals);
- $N_{feature}$ = 2–9;
- Random state_T = 0–10;

- Random state_M = 0–10.

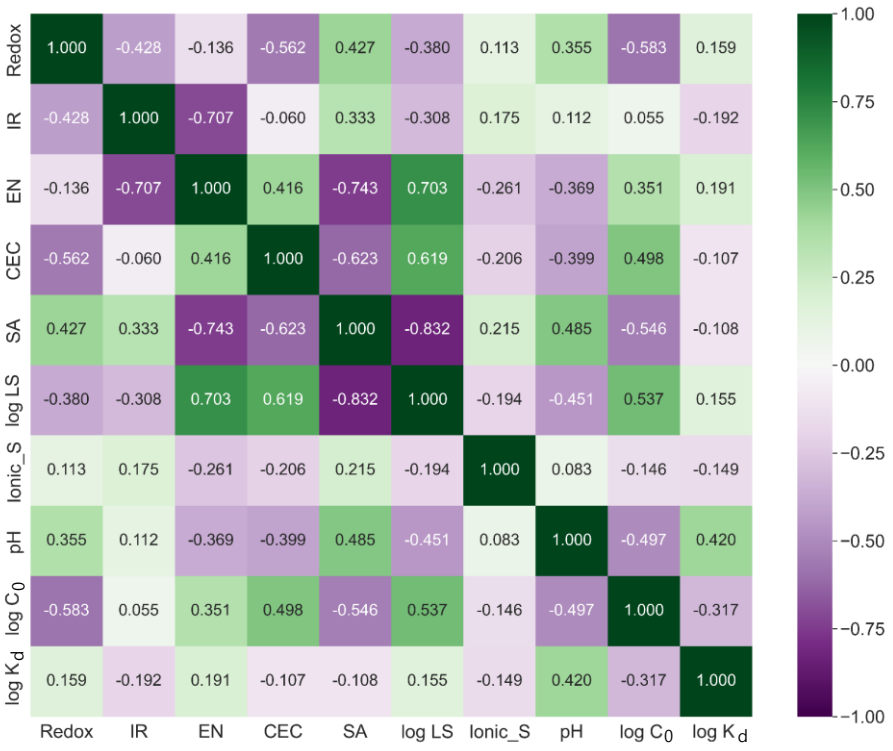

Figure 2. A Pearson correlation matrix for the various parameters calculated in the present work. IR: ionic radius; EN: electronegativity; CEC: cation exchange capacity; SA: surface area; Ionic_S: ionic strength; log LS: log of solid-liquid ratio; log C_0: log of initial radionuclide concentration; log K_d: log of K_d.

The other parameters were fixed at their default values [31]. Figure 3a presents the range of the R^2 value obtained by changing the Random state_T value. In the RF model, Random state_T is used as a random seed for the training and test sets classification, so a random change was observed in the R^2 result value with various Random state_T. According to the RF model calculation, the most optimal R^2 value was obtained at Random state_T = 8, and the lowest R^2 value was obtained at Random state_T = 4. In the present work, for a fixed Random state_T value, when the maximum R^2 value was high, the minimum R^2 value was also high, with a relatively small difference between the maximum and minimum R^2 values. Conversely, when the maximum R^2 value was low, the minimum R^2 value was also relatively low, with a significant difference between the maximum and minimum R^2 values. Based on the results obtained with the RF model, the maximum R^2 values derived from various calculation parameter ranges were determined to be R^2 = 0.9175, where RMSE = 0.3261.

On the other hand, when the RF model was calculated with the 5-5 nested cross-validation method, the deviations between the maximum and minimum values of R^2 derived by outer loops were relatively decreased (see Figure 3b). In addition, the R^2 values averaged from the outer fold calculation determined with various Random state_T were almost similar to each other. This may have been due to the reduction in the deviation of each result, caused by performing multiple cross-validations with five inner and five outer loops to avoid overfitting. According to the 5-5 nested cross-validation results, the most frequent $N_{feature}$ value, among the best $N_{feature}$ values obtained through the inner and outer fold calculation, was $N_{feature}$ = 2. In addition, the representative N_{tree} value was determined to be N_{tree} = 185 by averaging the best N_{tree} values and rounding to the nearest multiple of 5 for further calculation. Consequently, the final model test performed with the representative N_{tree} and $N_{feature}$ values, along with various Random state_M and Random state_T values, presented the averaged R^2 value of R^2 = 0.8604.

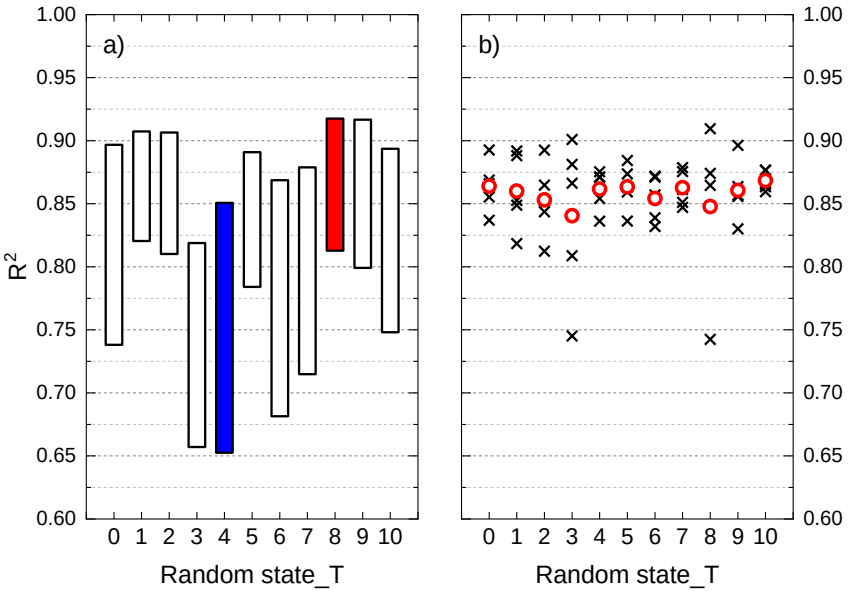

Figure 3. The R^2 values calculated with a fixed Random state_T using (**a**) the RF model and (**b**) the 5-5 nested cross-validation method, coupled with the RF model. Red and blue colored bars on the left-hand side represent the best and the worst results, respectively. Crosses on the right-hand side indicate the R^2 values obtained through five outer fold calculations at the given Random state_T, while circles represent the averaged R^2 value.

In order to assess the influence of the change of hyperparameters on the R^2 value, an additional evaluation was performed using the results representing the highest R^2 value among those shown in Figure 3a. Figure 4 presents the R^2 value according to the change in N_{tree} and $N_{feature}$ for the case of the general RF model. At this time, the Random state_T was fixed to Random state_T = 8 and the $N_{feature}$ was controlled in the range from 2 to 9. As a result, the highest R^2 value was obtained when N_{tree} = 105 for the general RF model with R^2 = 0.9175. As the N_{tree} value increased, the R^2 value changed slightly but converged to a constant value. This may have been due to the decrease in the diversity of the mean, as the number of samples of decision trees used for analysis increased with an increase in N_{tree}. In all cases, the R^2 value at N_{tree} = 1000 was relatively smaller than the maximum R^2 value.

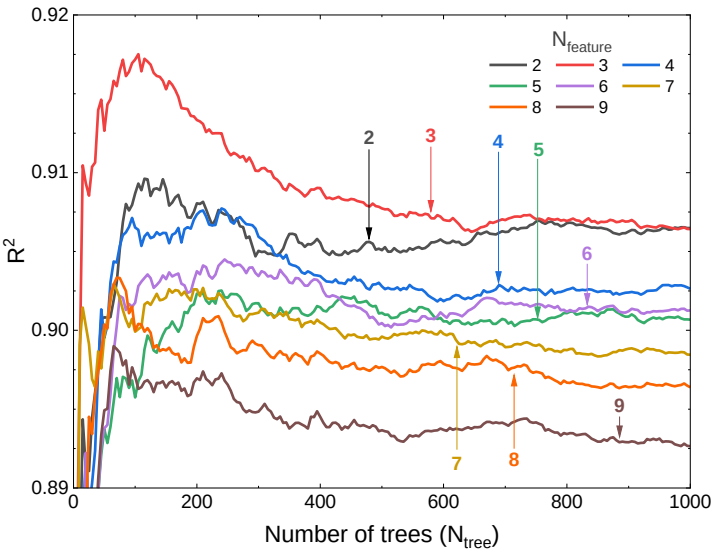

Figure 4. The values of R^2, calculated as a function of N_{tree} at Random state_T = 8, with the general RF model.

The $N_{feature}$ dependency of the R^2 value shown in Figure 4 will be discussed along with Figure 5, representing the change in R^2 value according to the Random state_M and $N_{feature}$ variables under two conditions of N_{tree} = 105 and N_{tree} = 1000 for the general RF model. As for two different values of N_{tree} = 105 and N_{tree} = 1000, the effects of the Random state_M and $N_{feature}$ variables on the distribution of R^2 values were considerably different. For N_{tree} = 105, the R^2 value was the highest at R^2 = 0.9175, but the distribution of relatively high R^2 values was somewhat limited only to the $N_{feature}$ = 3 and Random state_M = 5 points. For N_{tree} = 1000, the maximum R^2 value was R^2 = 0.9093, which was slightly low, but it still showed a broad distribution of relatively high R^2 values at low $N_{feature}$ values, regardless of the Random state_M variable. The machine learning results of the general RF model, based on the database constructed in this study, showed the maximum R^2 value at N_{tree} = 105, $N_{feature}$ = 3, Random state_T = 8, and Random state_M = 5. On the other hand, when it was set at N_{tree} = 1000 to prevent overfitting, the maximum R^2 value was observed at $N_{feature}$ = 2, Random state_T = 8, and Random state_M = 6.

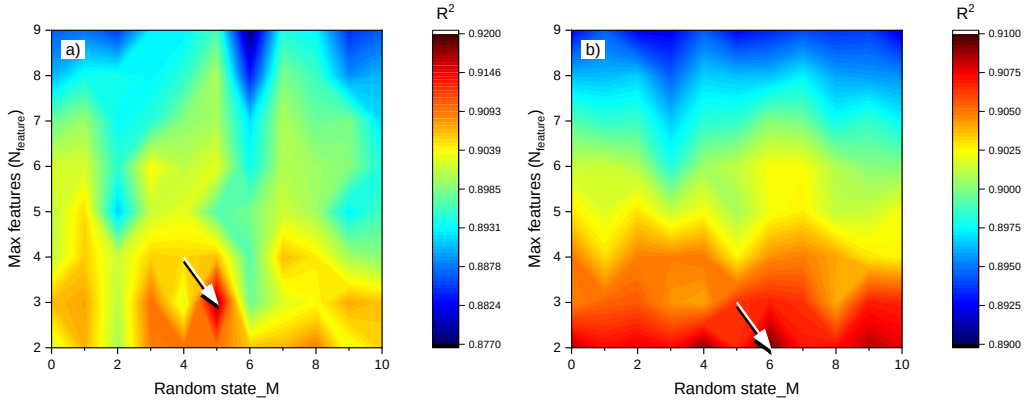

Figure 5. The R^2 mapping diagram, calculated with the general RF model for Random state_M and $N_{feature}$ at (**a**) N_{tree} = 35 and (**b**) N_{tree} = 1000. Random state_T was fixed to be Random state_T = 9 for both cases. The arrows point out the maximum R^2 values for each case.

Figure 5 suggests that the RF model based on the database constructed in this study seems helpful in calculating the best R^2 value by considering fewer input variables, i.e., a small $N_{feature}$, in dividing nodes. Typically, a relatively high R^2 value might be expected in the machine learning calculation with a large $N_{feature}$ value, as more input variables can be considered at once when calculating the model. However, when many input variables were considered at once in the present work, the nodes could not be subdivided and, thus, the diversity of decision trees was reduced, leading to a relatively low R^2 value [33–35].

Moreover, to examine the change in the maximum R^2 value, according to the number of data points used for the RF model training, 100, 200, 400, and 700 data points were randomly selected from the original database, with 777 data points being used to train the RF model. For the simple comparison, the ranges of the calculation parameters, in this case, were adjusted to be:

- Ntree = 10–500 (set in multiples of ten intervals);
- Nfeature = 2–9;
- Random state_T = 0–5;
- Random state_M = 0–5.

As shown in Table 1, with an increase in the number of data points used for machine learning, a remarkable increase in the maximum R^2 value and decrease in the RMSE value were observed, indicating that increasing the number of training data points would significantly improve the machine learning model result after calculation with hyperparameter and random seed adjustment.

Table 1. The maximum R^2 values obtained with various amounts of input data.

Amount of Input Data	Maximum R^2 Value	RMSE
100	0.7777	0.5807
200	0.7805	0.5520
400	0.8212	0.5522
700	0.8972	0.3543

4. Discussion

Figure 6a compares the experimental log K_d values with the predicted values derived from the RF model presenting the highest R^2 value, as established in the present work. For the comparison, the general RF model result that was calculated with N_{tree} = 105, $N_{feature}$ = 3, Random state_T = 8, and Random state_M = 5, displaying the maximum R^2 value of 0.9175, was used. Note that the training data for the RF calculation are not shown in Figure 6a. Consequently, a highly linear, 1:1 relationship between the predicted and experimental values of log K_d was observed, which indicates the usability of this machine learning method to predict the log K_d values for various radionuclides under arbitrary geochemical conditions with bentonite materials. In addition, it is expected that the accuracy of the log K_d prediction can be further improved by increasing the number of training data generally (see Table 1). Furthermore, the comparison between the experimental log K_d values and those predicted with the 5-5 nested cross-validation is shown in Figure 6b. According to the result, all five different fold loops from 1 to 5 show notable consistency with the overall R^2 value of 0.8683. Note that the data presented in Figure 6b only indicate the result determined by the five outer fold loops at Random state_T = 10 and are thus marginally different from the final model test result of R^2 = 0.8604.

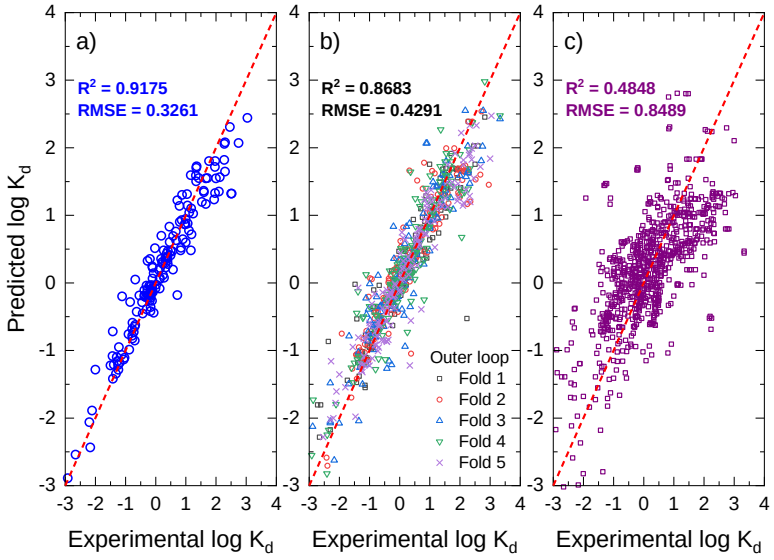

Figure 6. Comparisons of experimental log K_d value and predicted log K_d value, evaluated with (**a**) the general RF model presenting the highest R^2 value in this work, (**b**) the 5-5 nested cross-validation method with Random state_T = 10, and (**c**) the partial least-squares regression method.

Although the R^2 value obtained with the nested cross-validation method is slightly lower than that determined using the general RF model, the performance of the model established with the nested cross-validation method is still considered to be remarkable, based on the robustness and stability of the calculation model, together with the presumable mitigation of the overfitting problem.

In addition, Figure 6c presents the correlation between the experimental log K_d values and those estimated with the partial least-squares regression (PLSR) method [36,37]. The results calculated with the PLSR method show a relatively poor linear relationship between

the whole experimental log K_d data and the corresponding predicted ones, with a remarkably low coefficient of determination value of R^2 = 0.4848. The comparison among the results obtained with different approaches clearly indicates that the machine learning-based calculation outperforms the classical statistical method, i.e., PLSR. Note that the PLSR result shown in Figure 6c was obtained with the number of partial least-squares components set at 9.

Figure 7 shows the results of evaluating the relative importance of the input variables using the MDI method, with the RF model representing the highest R^2 value in this study. All nine input variables showed a certain level of contribution to predicting the log K_d value, according to the importance evaluation result. Most of all, the pH, followed by log C_0, EN, and LS, showed relatively higher criticality than the others.

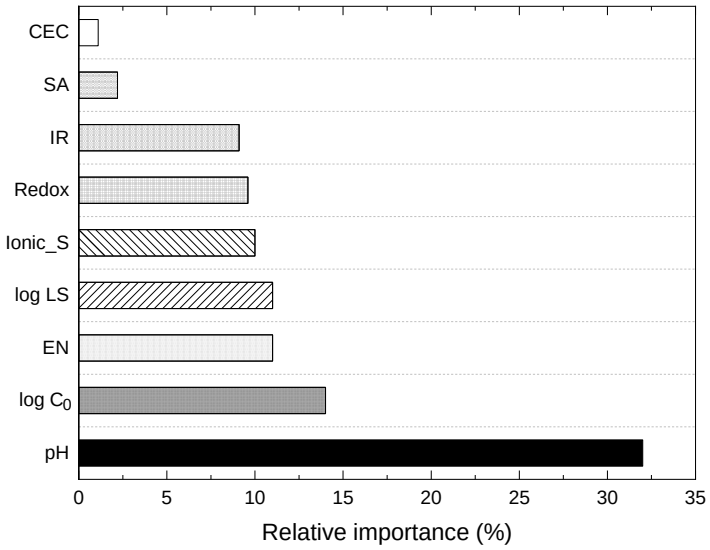

Figure 7. Relative importance of variables in terms of the estimation of log K_d values.

In particular, the pH is known to be a major variable having a sensitive effect on the surface charge of a mineral and on the hydrolysis reaction of radionuclides in the aqueous solution; hence, its importance can be explained and was quantitatively cross-checked via the RF model. Conversely, both CEC and SA, which are correlated to the mineralogical properties of bentonite, were found to have relatively low importance. It is assumed that the types of bentonite adopted in this study are not diverse, so their criticalities were evaluated to be insignificant, accordingly. The correlation between the input variables is expected to show a significant change when the input training data for assorted types of bentonites, with varied CEC and SA, are additionally considered in future investigations.

5. Conclusions

This study presented a model to predict the distribution coefficient, using the RF method and the experimental adsorption data taken from the JAEA-SDB. The database for the distribution coefficient prediction was constructed and the performance of the RF model results was evaluated, based on the R^2 and RMSE values. For the RF model calculation, various calculation parameters (i.e., N_{tree}, $N_{feature}$, Random state_T, and Random state_M) were controlled, and the relevant effects on the R^2 value were evaluated.

Previously, pH has been judged to be an important factor for the distribution coefficient, according to the various experimental results. The RF model and MDI method employed in the present work could quantitatively evaluate the importance of pH as a factor having a significantly higher contribution than the other input variables. In the present work, the R^2 values were determined to be R^2 = 0.9175 and R^2 = 0.8604 for the general RF model and the nested cross-validation method, respectively, suggesting that the use of the machine learning method in predicting the log K_d value would be notably meaningful.

This study is significant in that it has developed a model that can predict for the first time the distribution coefficient by applying a machine learning algorithm to three major bentonites that are used as backfill materials for deep-geologic repositories. The prediction model for the distribution coefficient, based on the RF model, is judged to be helpful in the preliminary estimation of the K_d values for unknown conditions. Furthermore, in addition to the 22 radionuclides trained in this work, it is expected to be used to predict the K_d value for the other radionuclides, based on the oxidation number, ionic radius, and electronegativity. The authors of this work note that the RF model obtained with a large N_{tree} value, i.e., N_{tree} = 1000, or the result obtained with the nested K-fold cross-validation method might be practically used to avoid the overfitting issue since the amount of absolute decrease in the R^2 value is not remarkably large.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/min12101207/s1, Table S1: Input variable database for the random forest model calculation [2,16–24].

Author Contributions: Conceptualization, J.-Y.L. and D.-H.K.; methodology, J.-Y.L. and D.-H.K.; software, D.-H.K.; validation, J.-Y.L.; formal analysis, J.-Y.L.; investigation, D.-H.K.; resources, J.-Y.L.; data curation, J.-Y.L. and D.-H.K.; writing—original draft preparation, D.-H.K.; writing—review and editing, J.-Y.L.; visualization, J.-Y.L.; supervision, J.-Y.L.; project administration, J.-Y.L.; funding acquisition, J.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea grant funded by the Korean government [No. NRF-2021M2E1A1085204], the Korean Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea [No. 20214000000410].

Data Availability Statement: The data used in this study are available within this article and supporting information.

Acknowledgments: The authors would like to thank Seonggyu Choi (KAERI) for his scientific advice and support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. SKB. *Design and Production of the KBS-3 Repository*; SKB Technical Report TR-10-12; SKB: Stockholm, Sweden, 2010.
2. Sugiura, Y.; Suyama, T.; Tachi, Y. *Development of JAEA Sorption Database (JAEA-SDB): Update of Sorption/QA Data in FY2019*; Japan Atomic Energy Agency: Ibaraki, Japan, 2020. [CrossRef]
3. Fernandes, M.M.; Vér, N.; Baeyens, B. Predicting the uptake of Cs, Co, Ni, Eu, Th and U on argillaceous rocks using sorption models for illite. *Appl. Geochem.* **2015**, *59*, 189–199. [CrossRef]
4. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
5. Zhang, W.; Li, Y.; Wu, C.; Li, H.; Goh, A.T.C.; Lin, H. Prediction of lining response for twin tunnels construction in anisotropic clays using machine learning techniques. *Undergr. Space* **2022**, *7*, 122–133. [CrossRef]
6. Nafouanti, M.B.; Li, J.; Mustapha, N.A.; Uwamungu, P.; AL-Alimi, D. Prediction on the Fluoride Contamination in Groundwater at the Datong Basin, Northern China: Comparison of Random Forest, Logistic Regression and Artificial Neural Network. *Appl. Geochem.* **2021**, *132*, 105054. [CrossRef]
7. Cipullo, S.; Nawar, S.; Mouazen, A.M.; Campo-Moreno, P.; Coulon, F. Predicting bioavailability change of complex chemical mixtures in contaminated soils using visible and near-infrared spectroscopy and random forest regression. *Sci. Rep.* **2019**, *9*, 4492. [CrossRef] [PubMed]
8. Jooshaki, M.; Nad, A.; Michaux, S. A Systematic Review on the Application of Machine Learning in Exploiting Mineralogical Data in Mining and Mineral Industry. *Minerals* **2021**, *11*, 816. [CrossRef]
9. Zhou, Z.H. *Machine Learning*; Tsinghua University Press: Beijing, China, 2016. [CrossRef]
10. Ray, S. A Quick Review of Machine Learning Algorithms. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 35–39. [CrossRef]
11. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
12. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* **1974**, *36*, 111–133. [CrossRef]

13. Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **2014**, *6*, 10. [CrossRef]
14. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef] [PubMed]
15. Louppe, G.; Wehenkel, L.; Sutera, A.; Geurts, P. Understanding variable importances in forests of randomized trees. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 431–439.
16. Chen, T.; Sedighi, M.; Jivkov, A.; Seetharam, S. A model for hydraulic conductivity of compacted bentonite–inclusion of microstructure effects under confined wetting. *Géotechnique* **2020**, *71*, 1071–1084. [CrossRef]
17. Physical and Chemical Data of Source Clays. Available online: https://www.clays.org/sourceclays_data/ (accessed on 1 May 2022).
18. Shannon, R.D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. Sect. A Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **1976**, *32*, 751–767. [CrossRef]
19. Little, E.J., Jr.; Jones, M.M. A complete table of electronegativities. *J. Chem. Educ.* **1960**, *37*, 231–233. [CrossRef]
20. Millero, F.J.; Feistel, R.; Wright, D.G.; McDougall, T.J. The composition of Standard Seawater and the definition of the Reference-Composition Salinity Scale. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **2008**, *55*, 50–72. [CrossRef]
21. Torstenfelt, B.; Andersson, K.; Allard, B. *Sorption of Sr and Cs on Rocks and Minerals*; National Council for Radioactive Waste: Stockholm, Sweden, 1981.
22. Torstenfelt, B.; Kipatsi, H.; Andersson, K.; Allard, B.; Olofsson, U. Transport of actinides through a bentonite backfill. In *Scientific Basis for Nuclear Waste Management-V*; Lutze, W., Ed.; Elsevier: New York, NY, USA, 1982; pp. 659–668. [CrossRef]
23. Mateus, M.V.; Araújo, L.S.; Leopoldino, A.B.; Ferreira, M.d.S.; Ferreira, D.C.; da Luz, M.S.; Gonçalves, J.C.S.I. Molecular interactions and modeling of anionic surfactant effect on oxygen transfer in a cylindrical reactor. *Environ. Eng. Sci.* **2019**, *36*, 180–185. [CrossRef]
24. Kitamura, A.; Tomura, T.; Sato, H.; Nakayama, M. *Sorption Behavior of Cesium onto Bentonite and Sedimentary Rocks in Saline Groundwaters*; Japan Atomic Energy Agency: Ibaraki, Japan, 2008.
25. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2, pp. 1–4. ISBN 978-3-642-00295-3.
26. Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295–316. [CrossRef]
27. Bernard, S.; Heutte, L.; Adam, S. Influence of hyperparameters on random forest accuracy. In *Multiple Classifier Systems*; Benediktsson, J.A., Kittler, J., Roli, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5519, pp. 171–180. [CrossRef]
28. Renaud, O.; Victoria-Feser, M.P. A robust coefficient of determination for regression. *J. Stat. Plan. Inference* **2010**, *140*, 1852–1862. [CrossRef]
29. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Sklearn.ensemble.RandomForestRegressor. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 1 May 2022).
32. Nested Cross-Validation. Available online: https://inria.github.io/scikit-learn-mooc/python_scripts/cross_validation_nested.html (accessed on 1 May 2022).
33. Gao, X.; Wang, L.; Yao, L. Porosity Prediction of Ceramic Matrix Composites Based on Random Forest. In Proceedings of the 3rd International Symposium on Application of Materials Science and Energy Materials (SAMSE 2019), Shanghai, China, 30–31 December 2019. [CrossRef]
34. Hou, N.; Zhang, X.; Zhang, W.; Wei, Y.; Jia, K.; Yao, Y.; Jiang, B.; Cheng, J. Estimation of surface downward shortwave radiation over China from Himawari-8 AHI data based on random Forest. *Remote Sens.* **2020**, *12*, 181. [CrossRef]
35. Shreyas, R.; Akshata, D.M.; Mahanand, B.S.; Shagun, B.; Abhishek, C.M. Predicting popularity of online articles using Random Forest regression. In Proceedings of the Second International Conference on Cognitive Computing and Information Processing, Mysuru, India, 12–13 August 2016. [CrossRef]
36. Rosipal, R.; Krämer, N. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques*; Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J., Eds.; Springer: New York, NY, USA, 2006; pp. 34–51. [CrossRef]
37. de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263. [CrossRef]