PREDICTING SOAKED CALIFORNIA BEARING RATIO OF A-2 LATERITIC SOIL USING SUPPORT VECTOR MACHINE AND RANDOM FOREST IN DELTA STATE

Okoro, V.U * and Ugbe F.C

Department of Geology, Faculty of Science, Delta State University, Abraka
*Corresponding author. victorokoro248@delsu.edu.ng

ABSTRACT

It normally takes four days or more in the laboratory to determine the soaked California bearing ratio (CBR) of lateritic soil samples. Furthermore, it is virtually impossible to conduct a large number of tests for a significant project within a short time frame. Alternative approaches, like forecasting models for soaked California bearing ratio, may therefore be used. The soaked California bearing ratio of A-2 lateritic soil was estimated using the Support Vector Machine (SVM) and Random Forest in western Niger Delta. A total of 52 dataset samples, comprising the plastic limit, liquid limit, plasticity index, percent of sand that passed through a 200-mesh sieve, moisture content, maximum dry density, and soaked CBR from a published source and laboratory results from a field investigation, were collected. Wakaito Environment for Knowledge (WEKA) 3.9.5 software was used to investigate the potential of SVMs and random forests to predict the soaked CBR of A-2 lateritic soil. It was found that support vector machine models outperformed random forest models in terms of estimating the soaked California bearing ratio of A-2 lateritic soil. The points on a scatter plot showing outputs from the training, cross-validation and percentage split 65% processes are very close to equality line.

Keywords: Index properties, soaked California bearing ratio, support vector machine and random forest

INTRODUCTION

Roads are necessary components of civil engineering infrastructure that link and enable access to other aspects of Man's social life in the society. The development of roads is dependent on the strength of the soil and the load that they must support over their lifetime for both vehicles and pedestrians. It will take a lot of time and money to sample along the proposed route in order to measure some of the key geotechnical indicators for a given area. The most widely used geotechnical property for measuring the overlay thickness of flexible pavements in Nigeria is the soil strength value known as California Bearing Ratio

(CBR), which can be calculated for both soaked and unsoakedsoil.California bearing ratio (CBR), is frequently used to indicate the strength of subgrades or subbases of pavement in relation to the strength of standard crushed rock samples (Yildirim and Gunaydin, 2011). When estimating the thickness (depth) of the subgrade or subbase layer of pavement for roads, railroads, and airports, the value of soaked and unsoakedCBR is employed (Taskiran, 2010). The CBR test is typically carried out on a soil sample that has been soaked or submerged in water to simulate the worst situation the subgrade material will be in during the construction of the pavement.The soil is typically soaked for 96 hours (4 days) during the CBR test, and if considerable loss is seen, the engineers may add an additional 5, 6, or 7 days of soaking to another sample in order to achieve the designed CBR. As a result, it is important to find a way to expedite CBR determination in order to reduce construction costs because CBR determination is a crucial part of construction projects (Yildirim and Gunaydin, 2011; Taskiran, 2010; Panagiotis et al., 2021). Additionally, if the CBR is calculated in a laboratory, the test findings might not be highly accurate because of sample disruption and preparation-related restrictions.Therefore, creating machine learning models is a quick and inexpensive way to estimate the CBR. KinMak (2006) and Taskiran (2010) claim that the CBR has also been connected to a number of the index aspects (Yildirim et.al. 2011). Black (1962) published a graph comparing the Plasticity Index (PI), Liquidity Index (LI), and CBR soil indicators for saturated clays.The correlation between CBR and suitability index, which depends on the plasticity and gradation of the soil, was made by Johnson and Bhatia in 1969. Agrawal and Ghanekar (1970) suggested the following connection in the form of an

equation: CBR is denoted by the formula 2.0-16.0*log (OMC)+0.07*LL (1), where OMC denotes the standard Proctor moisture content in fraction and LL denotes the liquid limit value of the soil. Also, a number of earlier research used both simple and multiple regression techniques to estimate the CBR in accordance with fundamental soil parameters (particle size, Atterberg's limit, MDD, OMC, etc.). Furthermore, these investigations generated a large number of regression equations. (Yildirim and Gunaydin, 2011; Alawi and Rajab, 2013; Erzin and Turkoz, 2016; Farias et al., 2018).However, it was found that none of the supplied formulae had any generalized solutions or high prediction accuracy, and the proposed regression equations in these studies were unable to generate a sufficient correlation (Yildirim and Gunaydin, 2011; Taskiran, 2010). This can be the result of intricate relationships between soil parameters and meaningless calculation methods.According to ASHTO classification, the dominant lateritic soil in Western Niger Delta is the A-2 type (Ugbe, 2011a).Therefore, it is essential to use machine learning techniques to forecast the soaked California bearing ratio of A-2 lateritic. In the current study, two machine learning algorithms were chosen to forecast A-2 lateritic soil's soaking CBR using physical soil indices. The 52 dataset samples comprised information from a published source (Ugbe, 2011b) on the plastic limit, liquid limit, plasticity index, percentage of particles, moisture content, maximum dry density, and soaked CBR, as well as laboratory findings from a field study as shown in figure 1. Weka 3.9.5 software is being used in the current study to create machine learning models to forecast soaked CBR of A-2 lateritic soil using support vector machine(SVM) and random forests in Western Niger Delta.

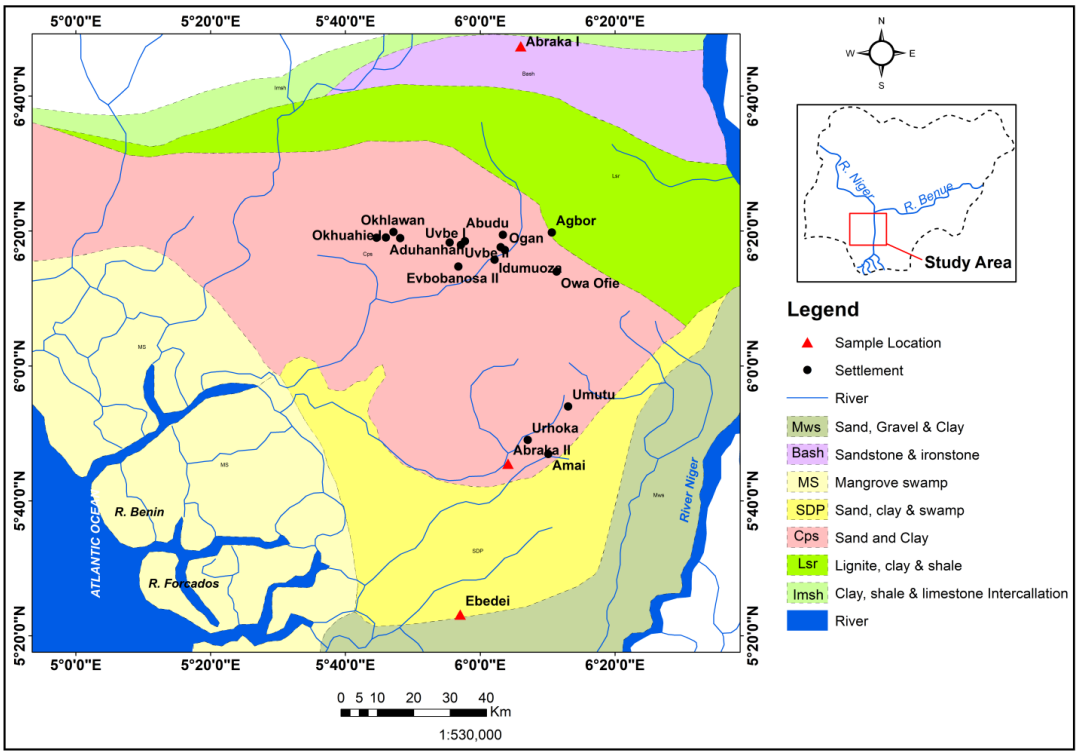

Figure 1: Map of the Location (Modified after Ugbe, 2011b)

MATERIALS AND METHODS

In this study, the following items were used: a GPS, field notebook, data sheets, sample bags, soil samples, sample labels, a trowel, and a spade.

Geotechnical Analysis of the Soils

A total of seven (7) samples from Abraka 1, Abraka II, and Ebedeiwere collected at a depth ranging from one (1) to seven (7) metres. The Omavic Geotechnical Laboratory in Warri, Delta State, Nigeria conducted the geotechnical investigation of the seven samples. According to BS 1377, classification tests as well as assessments of the moisture-density relationship and the soaked California bearing ratio were conducted. The following geotechnical tests were performed: fines, liquid limit, plastic limit, sand, maximum dry density, moisture

contents, and soaked California bearing ratio (C.B.R). In addition, 45 datasets, including %fines, liquid limit,plastic limit, %sand, maximum dry density, moisture contents, and soaked California bearing ratio of A-2 lateritic soil, were gathered from the published source (Ugbe, 2011).

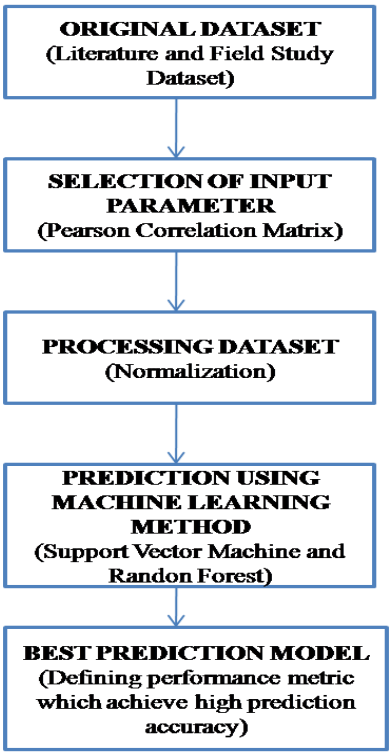

Figure 2: Flow Diagram of Prediction process

Model Development

Two (2) machine learning methods (ML) models—Support Vector Machine (SVM) and Random Forest (RF) models—were developed in order to predict the soaked California bearing ratio of A-2 lateritic soil in western Niger Delta (Figure 2). Each model was simulated using WEKA 3.9.5, the Waikato Environment for Knowledge Analysis (Figure 3).

Figure 3: Visual interface of Weka software

Support Vector Machine

$$y = f(x) = w\theta(x) + b$$

Support vector machine (SVM), a popular machine learning technique that was first introduced by Vapnik in 1999, is frequently used to handle a variety of real-world issues, including the prediction of soil-related variables. The central idea of SVM is to use a hyperplane to map the original input space into a high-dimensional feature space (Bui et al, 2016). Let x = xi be a collection of input variables utilized in the models, and let y be the result (predicted variable). Equation represents the SVM function.

If b denotes the model's bias, w is its weight matrix, and (x) is the term used to describe a feature that is nonlinearly mapped from the input space x.

Random Forests

A tool for classification and regression is the Random Forest. Many tree predictors are used in this strategy. In this method, a random vector was chosen at random from the input vector and used to build each tree (Figure 4). Instead of the classification labels used by the RF classifier, the Tree

predictor makes use of numerical values. The Random Forest Regression (RFR) method builds a tree by using a combination of parameters or a single parameter (chosen at random) at each node. A strategy for creating training data called bagging involves replacing randomly chosen data with data from the original data set aside for training. For each feature combination, the training data can also be randomly selected to produce a unique tree. In the bagging procedure, 35% of the original data were excluded from each tree formed while 65% of the original data were utilized for training. A pruning mechanism and a variable selection process were required to build a tree predictor. RFR used the first method to select the study's variable measure.TheGini index approach establishes the impurity of the variable in respect to the result.RFR permits the tree to grow to the maximum depth of the training data by merging factors, and fully grown trees are not permitted to be pruned back.As a result, the RFR has an advantage over the M5P.

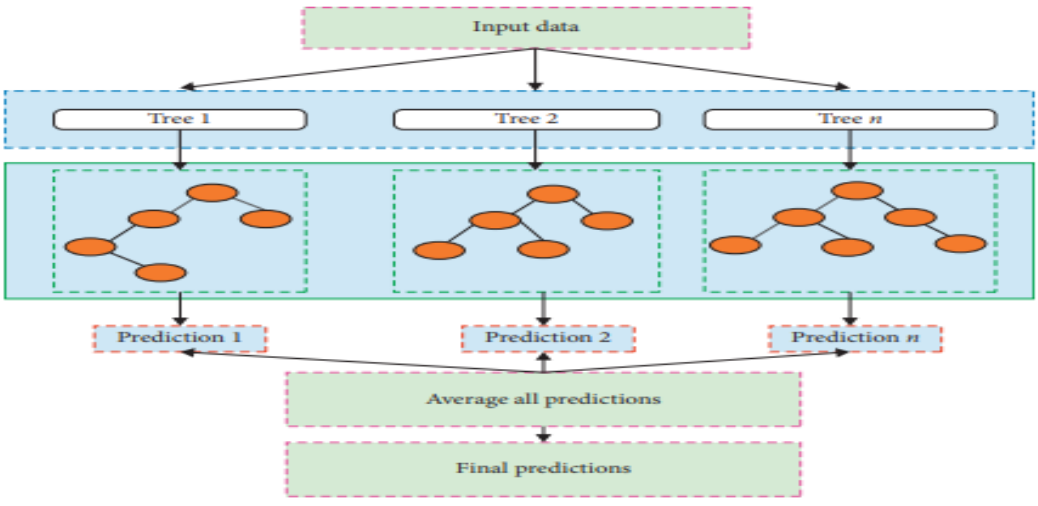

Figure 4: The Random Forest Model Structure (Hayder et al, 2021)

Selection of Input Parameter

A technique must be used successfully, and this depends on choosing the right input parameters. The input parameters (sand%,% fines, Liquid limit LL, Plastic limit PL, Plasticity Index PI, Optimal moisture Content OMC, and Maximum Dry Density MDD) were chosen based on the Pearson correlation coefficient (R) with the output (soaked California Bearing Ratio). The less significant the correlation, the higher the linear correlation is between the tested variable and the absolute value of the correlation coefficient, which is closest to the value of 1 (Akinwamide et al, 2022). More than 0.81, 0.61-0.80, 0.41-0.60, 0.21-0.40, and less than 0.2 of the coefficient of correlation, respectively, indicate very strong, strong, moderate, and no association.Using Microsoft Excel 2019, the Pearson's correlation coefficient for the datasets was computed.

Training and Testing

Using models from support vector machines (SVM), and random forests (RF), the WEKA 3.9.5 was trained and tested via the Classify module. Cross validation and percentage split testing were the testing techniques used. A predetermined percentage split was employed, with 65% (34 dataset) for training and the remaining 35% (18 dataset) for testing. Cross validation testing mode used a 10-fold split, which means that the data was divided into ten (10) equal parts, nine (9) of which were utilized for training and one (1) for testing the model.

Performance Evaluation Metrics for Classifiers

Performance metrics like the Coefficient of Determination, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error were used to assess the precision of the Classify using Support Vector Machine (SVM) and Random Forest (RF) models

Coefficient of Determination (R^2 or R squared)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \quad (1)$$

(worst value = $-\infty$; the best value is +1) (Sorensen and Okkels, 2013).

Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{m}}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (2)$$

(best value =0, worst value = $+\infty$

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \quad (3)$$

(best value= 0, worst value = $+\infty$)

Relative Absolute Error

$$\text{RAE} = \frac{\sum_{i=1}^{n}|\text{CBR}_{mi} - \text{CBR}_{pi}}{\sum_{i=1}^{n}|\text{CBR}_{mi} - \text{CBR}_{m}} \quad (4)$$

The relative absolute error (RAE) has a simple meaning: if the RAE is less than one, the model outperforms the fundamental model. A perfect model has a relative absolute error of 0. Ideal RAE should be as close to zero as is practical.

RESULTS AND DISCUSSIONS

Table 1 provides statistical descriptions of the soil properties that were evaluated. From Table 1, it is clear that the distribution of the median and mean values for the soil parameters is quite uniform. This demonstrates that the results from soil experiments are roughly regularly distributed.

Table 1: Descriptive Statistics of datasets used in this study

Description	%Fines	L.L%	P.L %	P.I %	%Sand	MDD Kg/m3	OMC %	Soaked CBR Value%
Minimum	14	22.2	15	4.6	58	1734	7.7	3
Maximum	42	46.5	30	26	86	2120	14	43
Mean	27.84615	34.84423	21.20962	13.78846	72.21154	2011	10.36538	18.80769
Mode	31	33.5	20	16	69	2040	10	13
Median	27.5	34.55	20	14	73	2030	10	16
Standard Deviation	6.322647	6.697539	3.634058	4.998967	6.408746	77.35683	1.610978	9.976518
Variance	39.97587	44.85702	13.20638	24.98967	41.07202	5984.078	2.595249	99.53092

Table 2: Pearson's Correlation Coefficient (r) for the datasets

	%Fines	L.L %	P.L %	P.I %	%SAND	MDD Kg/m3	OMC %	SOAKED CBR VALUE%
%Fines	1							
L.L %	0.645174	1						
P.L %	0.70077	0.657453	1					
P.I %	0.377624	0.857405	0.188459	1				
%SAND	-0.99167	-0.65132	-0.69071	-0.38612	1			
MDD Kg/m3	-0.30733	-0.09851	-0.1753	-0.03103	0.29691	1		
OMC %	0.601237	0.376488	0.416036	0.231035	-0.58195	-0.26182	1	
SOAKED CBR VALUE%	-0.41982	-0.33502	-0.23916	-0.23889	0.440727	0.032521	-0.22015	1

Table 2 contains the correlation matrix. A strong link between liquid limit and plasticity index was shown by the matrix. In order to prevent multicollinearity, plasticity index were removed from the models (Iyeke et al, 2016).To reduce bias in the machine learning algorithms for one feature over another, a system designer ideally wants the same range of values for each input feature (Phani et al, 2015). By starting the training process for each feature on the same scale, data normalization can help save training time(Table 3).

Table 3: Descriptive Statistics of Normalized datasets

Description	L.L%	P.L %	%Sand	MDD Kg/m3	OMC %	%Fines	Soaked CBR Value%
Minimum	-1	-1	-1	-1	-1	-1	14
Maximum	1	1	1	1	1	1	42
Mean	0.041	-0.172	0.015	0.435	-0.154	-0.21	27.846
Standard Deviation	0.551	0.485	0.458	0.401	0.511	0.499	6.323
Variance	0.303601	0.235225	0.209764	0.160801	0.261121	0.249001	39.980329

Scatter plot of Actual soaked CBR against Predicted soaked CBR during Training (A), Cross Validation (k-10) (B) and percentage split(65%) (C) using Support Vector Machine and Random Forest are shown in figure 5 and 6.

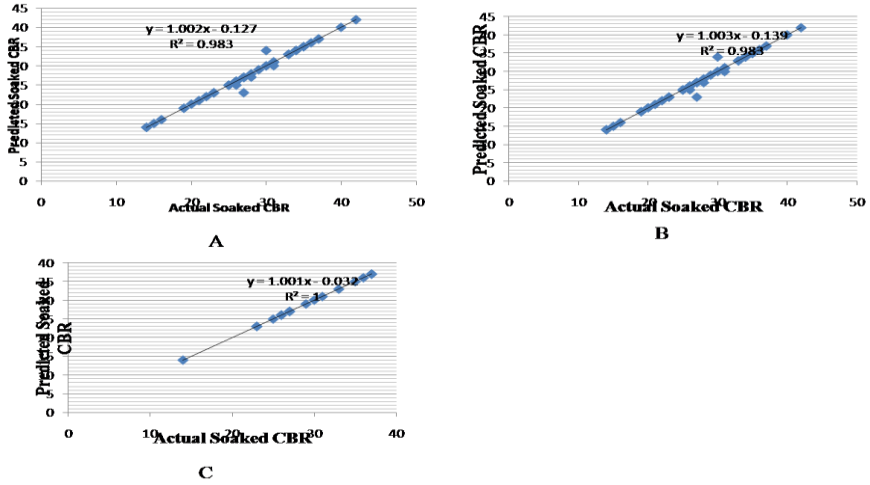

Figure 5: Scatter plot of Actual soaked CBR against Predicted soaked CBR during Training (A), Cross Validation (k-10) (B) and percentage split (65%) (C) using support vector machine

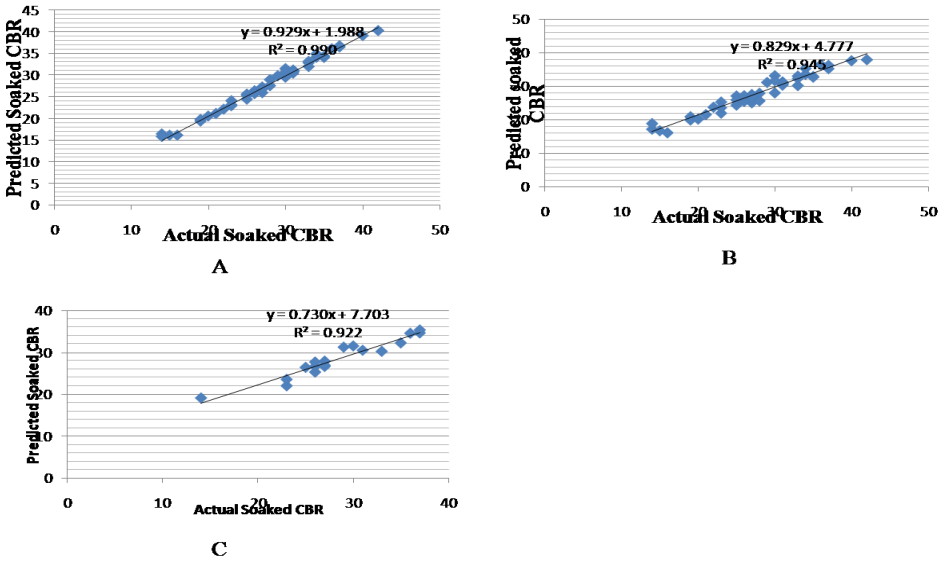

Figure 6: Scatter plot of Actual soaked CBR against Predicted soaked CBR during Training (A), Cross Validation (k-10) (B) and percentage split (65%) (C) using random forest model

Table 4: Comparison of Machine Learning Model using Performance Metrics

Training/	Machine Learning Model	(R^2)	MAE	RMSE	RAE (%)
Training dataset	Support Vector Machine	0.983	0.2233	0.8139	4.5362
	Random Forest	0.990	0.5073	0.7159	10.3047
Cross validation (K-10)	Support Vector Machine	0.983	0.234	0.8222	4.7335
	Random Forest	0.945	1.2144	1.6444	24.5624
Percentage split (65%)	Support Vector Machine	1	0.0195	0.0245	0.4447
	Random Forest	0.922	1.5428	1.9457	35.1257

Table 4compared between the actual soaked and predicted soaked CBR for training, cross validation, and percentage split (65%) using support vector machine and random forest models. As observed, the predicted outcomes came close to the

expected values. The R^2values of 0.983, 0.983, and 1 correspondingly show how well the training, cross validation, and percentage split (65%) predicted the outcome (Sorensen and Okkels, 2013). For training, cross validation, and percentage split of 65%, the RMSE values were 0.8139, 0.8222, and 0.0245, respectively; MAE values were 0.2233, 0.234, and 0.0195 for training, cross validation, and percentage split(65%) (Erzin et al, 2010); RAE values were 4.5362, 4.7335, and 0.4447 for training, cross validation, and the 65% split (Pham, et al 2022), respectively using support vector machine models. Table 4 also showed the comparison between the actual soaked and predicted soaked CBR for training, cross validation, and percentage split (65%) using random forest. The training, cross validation, and percentage split (65%) all produced close predictions, as indicated by the R^2 values of 0.990, 0.945, and 0.922, respectively (Sorensen and Okkels, 2013).For training, cross validation, and a 65% split, the RMSE values are 0.7159, 1.6444, and 1.9457 respectively.TheMAE values for training, cross validation, and percentage split (65%) are 0.5073, 1.2144, and 1.5428 respectively, illustrating the evaluation of the actual and predicted results(Erzin et al, 2010). The relevant RAE values for training, cross validation, and the percentage split 65% were 10.3047, 24.5624, and 35.1257 (Pham, et al 2022).

In addition to the seven (7) additional laboratory results from the intensive field study, 45 datasets of A-2 lateritic soil were collected from the published articles (Ugbe, 2011b). Additionally, this dataset includes the following variables: % fines, % liquid limit, % plastic limit, % plasticity index, % sand, % maximum dry density, % optimum moisture content, % soaked california bearing ratio. The normalization procedure were used to the dataset rescaling(Phani et

al, 2015).The input (% fines, % liquid limit, % plastic limit, % plasticity index,%sand, % maximum dry density, % optimum moisture content) and output (soaked californiabearing ratio) of A-2 lateritic soil was normalized using the min-max and log normalization functions.The relationship between the characteristics of the granular size distribution, on the other hand, has the highest association with soaked CBR.According to the coefficient of determination (r^2), root mean square error (RMSE), root absolute error (RAE) and mean absolute error(MAE),the performance of built-in models has been calculated.

When comparing machine learning models using performance metrics, it was discovered that support vector machines performed better at predicting the soaked California bearing ratio of A-2 lateritic soil than the random forest.

CONCLUSION

In this study, support vector machine and random forest models were built and trained to predict soils' soaking CBR. The input parameters for the models were the plastic limit, liquid limit, fineness percentage, sand percentage, moisture content, and maximum dry density.The constructed support vector machine and Random Forest model were reliable models in forecasting soaked CBR based on the obtained R^2 value for training, cross validation, and percentage split (65%).Also, it was discovered that support vector machines outperformed random forest machine learning methods for forecasting the soaked California bearing ratio of A-2 lateritic soil since the support vector model had lower values of RMSE, MAE, and RAE.

ACKNOWLEDGMENT

Sincerely grateful to my supervisor, Dr. F.C. Ugbe, for allowing me to utilize his data for this study.

REFERENCES

Agarwal KB and Ghanekar KD.,1970, Prediction of CBR from plasticity characteristics of soil.In: Proceeding of 2nd south-east Asian conference on soil engineering , Singapore . Bangkok: Asian Institute of Technology; p.571–6

Akinwamide J.T, Jacob O.E, Osuji S.O and Ebuka. N (2022) Application of soft computing techniques in modeling soaked and unsoaked California bearing ratio. Asian Soil Research Journal, 6:32- 46.

Alawi, M.H., and Rajab, M.I, 2013. Prediction of California bearing ratio of subbaselayerusing multiple linear regression models. Road Materials and Pavement Design, 14 (1), 211–219.

Black WPM.1962, A method of estimating the CBR of cohesive soils from plasticity data. Geotechnique Journal;12:271–272

Black, M., Model sand Metaphors Studies in Language and Philosophy. Madrid Cornell University Press, 1962. BS1377 1990. Methods of Testing Soil for Civil Engineering Purposes. British Standards Institute, London.

Bui D.T, Tuan, T.A, Klempe, H., Pradhan, B and Revhaug I.(2016) " Spatial prediction model computing techniques in modeling soaked and unsoaked California bearing ratio. Asian Soil Research Journal, 6:32- 46.

Erzin, Y., and Turkoz, D, 2016. Use of neural networks for the prediction of the CBR value of some Aegean sands. Neural Computing and Applications, 27 (5):1415–1426.

Erzin, Y., Hanumantha Rao, B., Patel, A., Gumaste, S.D. and Singh, D.N. 2010. Artificial neural network models for predicting electrical resistivity of soils from their thermal resistivity. International Journal of Thermal Sciences , 49:118–130.

Gunaydin, O., Gokoglu, A., and Fener, M., 2010 "Prediction of artificial soil's unconfined compression strength test using statistical analyses and artificial neural networks", Advances in Engineering Software, 41:1115–1123.

HayderR.,Mohammed M. **and** Sumarni Ismail(2021)Random Forest versus Support Vector Machine Models' Applicability for Predicting Beam Shear Strength. Hindawi journal, 2021:

Iyeke S. D, Eze E. O, Ehiorobo J. O and Osuji S. O (2016 Estimation of Shear Strength Parameters of Lateritic Soils Using Artificial Neural Network. Nigerian Journal of Technology (NIJOTECH),35(2) : 260 – 269

Johnson D. G., Bhatia H. S (1969)"The engineering characteristics of the lateritic gravels of Ghana". Proceedings of 7th International Conference on Soil Mechanics and Foundation Engineering, Mexico August 28 29. Bangkok: Asian Institute of Technology. 2:13 – 43.

Kin Mak Wai.,2006, California bearing ratio correlation with soil index properties master of engineering (civil–geotechnics), Faculty of Civil Engineering,UniversityTeknology Malaysia;

Panagiotis G. A., Athanasia D. S., Abidhan B., Pijush S., Kypros P. (2021)Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models.Cement and Concrete Research Journal, 145**:** 106449

Phani K.V., Manjula C.H., and Poornima P. (2015) Artificial Neural Networks (ANNS)

For Prediction of California Bearing Ratio of Soils, International Journal Of Modern Engineering Research (IJMER), 5: 2249–6645

Pham Q. B., Kumar M and Nunno ., F. D (2022), "Groundwater level prediction using machine learning algorithms in a drought-prone area," Neural Computing & Applications journal 34(13):10751–10773

Sorensen, K.K. and Okkels, N. (2013), Correlation between drained shear strength and plasticity index of undisturbed over consolidated clays. Preceedings of the 18th international conference on soil mechanics and Geotechnical Engineering, Paris, pp. 1-6

Taskiran,T., 2010 "Prediction of California bearing ratio (CBR) of fine grained soils by AI methods", Advances in Engineering Software, 41:886–892.

Ugbe, F.C. (2011a) Estimating compaction cycles characteristics from fines in A-2 type lateritic soil. Research Journal of Environmental and Earth Sciences, 3(4): 433-437

Ugbe F.C (2011b) Basic Engineering Geological Properties of Lateritic Soils from Western Niger Delta, Research Journal of Environmental and Earth Sciences, 3(5): 571-577

Vapnik V. N.(1998) "Statistical learning theory," New York: Wiley.

Yildirim, B., and Gunaydin, O., 2011"Estimation of California bearing ratio by using soft computing systems", Expert Systems With Applications, 38:6381-6391.