East African Journal of Engineering
eaje.eanso.org
Volume 8, Issue 1, 2025
Print ISSN: 2707-5389 | Online ISSN: 2707-5397
Title DOI: https://doi.org/10.37284/2707-5397

EAST AFRICAN NATURE & SCIENCE ORGANIZATION

Original Article

Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya

Billy Kipchirchir Koech[*], *Dr. Simpson Nyambane Osano, PhD*[1] *& Dr. Abraham Mutunga Nyete, PhD*[1]

[1] University of Nairobi, P. O. Box 30197-00100, Nairobi, Kenya.
[*] Author for Correspondence ORCID ID; https://orcid.org/0009-0007-1164-0547; Email: billykoech@students.uonbi.ac.ke

Article DOI : https://doi.org/10.37284/eaje.8.1.2651

Publication Date:

29 January 2025

Keywords:

California Bearing Ratio (CBR), Artificial Intelligence (AI), Machine Learning (ML), Soil Index Properties, Extreme Gradient Boosting (XGBoost)

ABSTRACT

The California Bearing Ratio (CBR) is an important civil and transportation engineering test. It is normally carried out to assess soil's bearing capacity and strength for road pavement and foundation construction. The test, however, is both time-consuming and labour-intensive, resulting in significant delays during the construction process, ultimately leading to financial losses due to the high cost typically associated with construction projects. As a potential solution to this issue, an investigation is conducted into the application of artificial intelligence (AI) and machine learning (ML) techniques for accurately forecasting CBR values. Three models were used in the study, namely, the random forest model, linear regression model, and extreme gradient boosting (XGBoost) model. These models were employed to forecast CBR values based on several soil index properties. These properties included particle size distribution (i.e., percentage of soil passing through the sieve of diameter 0.425mm and 0.075mm), liquid limit (LL), plasticity index (PI), maximum dry density (MDD), plastic limit (PL), and optimum moisture content (OMC). A dataset containing these soil properties and corresponding CBR values for soils was obtained from the University of Nairobi civil engineering laboratory. The models were then trained on 80% of the data and tested on 20%. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of determination (R^2) were used to evaluate the accuracy of the predictions. The findings showed that XGBoost was the most accurate model with the lowest MAE, MSE, and RMSE, and the highest R^2, making it the preferred model for predicting CBR.

APA CITATION

Koech, B. K., Osano, S. N. & Nyete, A. M. (2025). Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya. *East African Journal of Engineering*, *8*(1), 46-60. https://doi.org/10.37284/eaje.8.1.2651

CHICAGO CITATION

Koech, Billy Kipchirchir, Simpson Nyambane Osano and Abraham Mutunga Nyete. 2025. "Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya". *East African Journal of Engineering* 8 (1), 46-60. https://doi.org/10.37284/eaje.8.1.2651.

HARVARD CITATION

Koech, B. K., Osano, S. N. & Nyete, A. M. (2025) “Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya”, *East African Journal of Engineering*, 8(1), pp. 46-60. doi: 10.37284/eaje.8.1.2651.

IEEE CITATION

B. K., Koech, S. N., Osano & A.M., Nyete “Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya” *EAJE*, vol. 8, no. 1, pp 46-60, Jan. 2025.

MLA CITATION

Koech, Billy Kipchirchir, Simpson Nyambane Osano & Abraham Mutunga Nyete. “Application of Machine Learning in Estimating California Bearing Ratio from Soil Index Properties in Kenya” *East African Journal of Engineering*, Vol. 8, no. 1, Jan. 2025, pp. 46-60, doi:10.37284/eaje.8.1.2651.

INTRODUCTION

The California Bearing Ratio (CBR) is an important and widely recognized test in civil and transportation engineering. Specifically, it is essential for determining the suitability of soil for various engineering applications, such as road pavement and foundation design (ASTM D1883-16, 2016). This test plays a pivotal role in evaluating the strength and bearing capacity of subgrades, subbase, and base course materials, which are crucial elements in road construction.

The CBR test was originally developed in the 1930s for the California Division of Highways by O. J. Porter. Its primary purpose was to facilitate the evaluation and classification of soil for engineering applications (Yoder et al., 1975). The test involves driving a standard-diameter plunger, at a controlled rate of 1mm/min into a compacted soil sample. Before this, the soil sample is soaked for four days, simulating a worst-case scenario, such as continuous rainfall over an extended period. This step ensures that the results reflect the soil’s performance under extreme moisture conditions.

During the test, the load needed to achieve specific penetrations is carefully measured, and the CBR value is calculated as the ratio of the applied load to the load required to penetrate standard crushed rock. The corresponding values for crushed rock penetration are shown in Table 1 (Nguyen, & Mohajerani, 2015). While the CBR test is straightforward and relatively inexpensive to conduct, it has a significant disadvantage: the time-consuming process of soaking the soil sample for four days. This prolonged testing period often leads to delays in construction projects, which can be costly.

Table 1: Load Penetration for Standard Crushed Rock with Cbr = 100%

Penetration depth (mm)	Load (kN)
2	11.5
2.5	13.24
4	17.6
5	19.96
6	22.2
8	26.3
10	30.3
12	33.5

One potential solution to mitigate this challenge is the application of artificial intelligence (AI) and machine learning (ML) techniques for predicting CBR values. When trained on large datasets, machine learning algorithms can detect patterns and deliver highly accurate predictions (Janiesch et al., 2021). The roots of ML trace back to the mid-20th century when the concept of AI first began to emerge and gain traction. Pioneering researchers like Alan Turing and Claude Shannon made significant contributions to the field. However, it was in the 1950s and 1960s that the foundations of ML were established (Çelik, 2018).

A significant early milestone in machine learning was the creation of the perceptron by Frank Rosenblatt in 1957. The perceptron, a type of artificial neural network, demonstrated the ability to learn and make predictions, laying the groundwork for further advancements in neural networks and deep learning (Keith D. Foote, 2019). The modern era of machine learning began in the 2000s. These progressions were largely driven by the widespread availability of digital data and significant improvements in computational power. This period also saw the emergence of "big data" and the creation of sophisticated algorithms capable of analyzing vast datasets (Firican, 2022).

ML is generally categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training models using labelled datasets to recognize patterns and make precise predictions. In contrast, unsupervised learning utilizes unlabeled datasets, enabling models to discover hidden patterns or structures within the data, while reinforcement learning relies on a reward-based system to train models to make decisions that optimize desired outcomes (Sasakawa et al., 2008).

This paper focuses on supervised learning methods, utilizing regression techniques to predict CBR values. Regression, a statistical approach, explores the relationship between dependent and independent variables. This study utilized regression analysis to explore the relationship between soil index properties and corresponding CBR values, forming the basis for predictive modelling.

METHODOLOGY

Dataset

Training good models require large amounts of data, which are often gathered from various sources to ensure diversity and representativeness. This will enable models to efficiently learn the intricate connection between the dependent and independent variables. By exposing the models to vast amounts of well-structured data, they gain the capability to discern patterns within the dataset, therefore significantly enhancing their learning capacity and accuracy in predictions.

The data used in this research was meticulously obtained from the Civil Engineering Department laboratory at the University of Nairobi. This dataset was carefully selected for several compelling reasons. First and foremost, the experiments were carried out under controlled conditions that can be precisely replicated, ensuring reliability and consistency in the research results. Secondly, the diversity of soils examined in the tests is extensive, encompassing a wide range of soil types prevalent across Kenya’s diverse regions. Lastly, the dataset within the laboratory archives was deemed sufficient enough to support the comprehensive development and validation of machine learning models for this study.

Soil Index Properties (Input Parameters)

Several soil index properties were taken into account in this research. They include particle size distribution (percentage of soil passing through 0.425mm and 0.075mm diameter sieves), liquid limit (LL), plastic limit (PL), plasticity index (PI), maximum dry density (MDD), and optimum moisture content (OMC). Table 2 shows the statistical parameters of the data.

Table 2: Statistical Parameter of The Dataset

	LL	PL	PI	MDD	OMC	0.425mm	0.075mm
count	519	519	519	519	519	519	519
mean	53.94	28.77	25.30	1563.05	23.03	67.99	56.97
std	13.97	7.72	8.79	235.10	7.58	24.35	25.14
min	14.00	9.00	2.00	1160.00	6.70	8.00	5.00
25%	43.00	23.00	20.00	1373.50	17.15	49.00	36.00
50%	53.00	29.00	25.00	1532.00	23.10	75.00	58.00
75%	64.00	34.00	31.00	1737.00	28.30	89.00	80.50
max	86.00	51.00	51.00	2220.00	45.80	99.00	99.00

Methods Used

In this research, the Python programming language was used alongside the Jupyter Notebook environment for data processing and visualization. An initial exploratory data analysis (EDA) was carefully conducted to gain a clearer and more detailed insight into the underlying structure of the data. A total of 80% of the data was used for model training and 20% was allocated for precise model testing. Three different models, random forest, linear regression, and extreme gradient boosting, were strategically selected to predict the California bearing ratio (CBR) values. To assess the accuracy and robustness of the models, various well-established metrics were utilized, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

Exploratory Data Analysis (EDA)

EDA is important because it provides a deeper understanding of the dataset before diving into modelling and ensures that potential issues are addressed early. It helps uncover hidden patterns, trends, and relationships in the data. EDA also plays a crucial role in identifying data quality issues, such as inconsistencies or errors, and selecting appropriate modelling techniques (Sylvia, & Murphy, 2023).

The dataset had no missing values or anomalies. Density plots of the different independent variables were plotted as shown in Figures 1 (a)-(g). Density plots are used to visually represent the distribution of the data and help identify any deviations from normality.

Figure 1(a); Liquid limit distribution

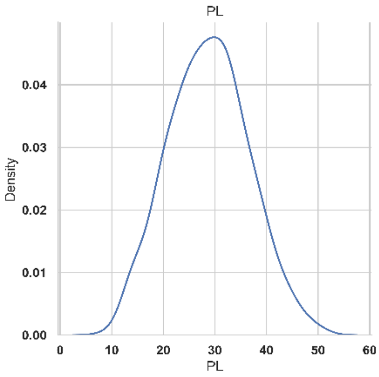


Figure 1(b): Plastic Limit Distribution

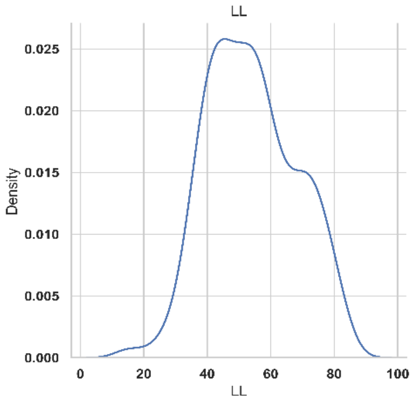


Figure 1(c): Plasticity Index Distribution

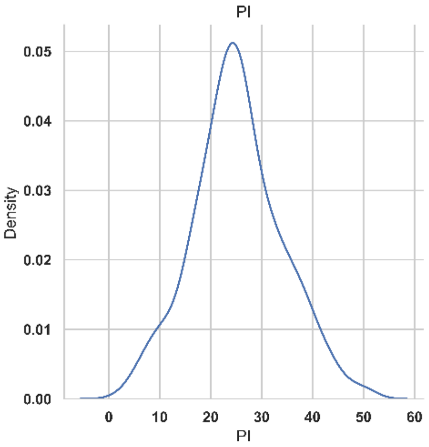


Figure 1(d): Maximum dry Density Distribution

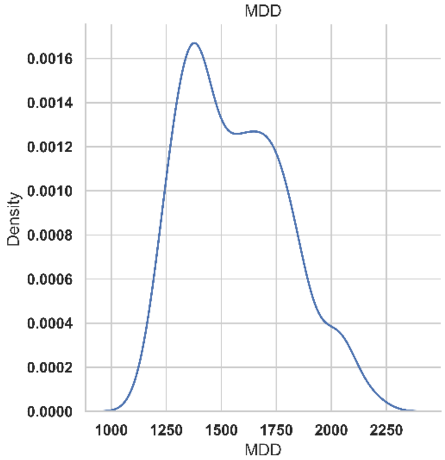


Figure 1(e): Optimum Moisture Content Distribution

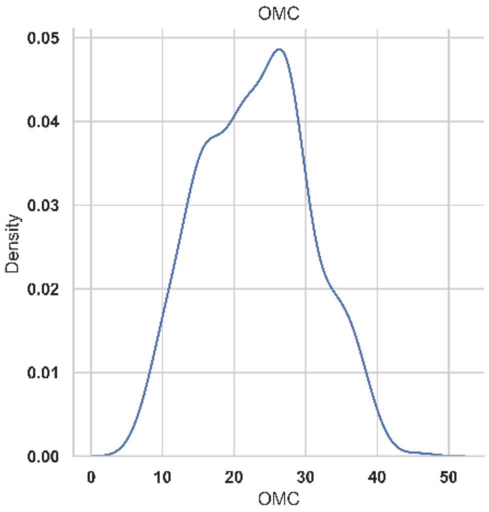


Figure 1(f): Distribution of the Percentage of Soil Particles Passing through Sieve Size 0.425mm

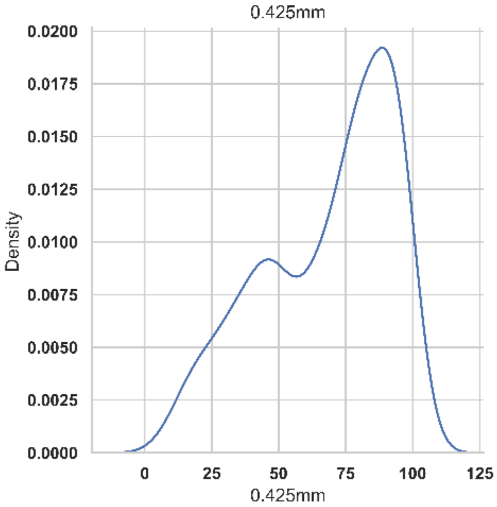
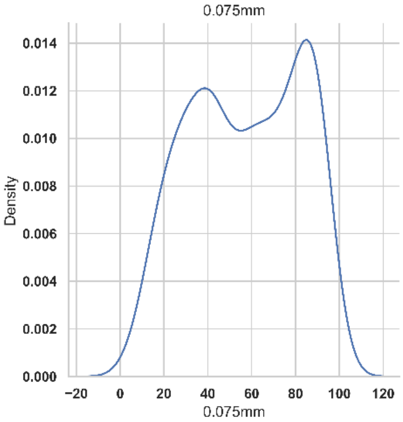


Figure 1(g): Distribution of the Percentage of Soil Particles Passing through Sieve Size 0.075mm



From the figures above, the majority of the variables are normally distributed, suggesting that the dataset was well-prepared and consistent. Therefore, no manipulation was done on the data before modelling, ensuring the integrity of the original dataset.

Figure 2 shows the distribution of CBR values in the dataset. Figure 2(a) represents the distribution of original CBR values, which were right-skewed, with the mean value being significantly greater than the median. To achieve a normal distribution, the original CBR values were transformed using the natural logarithm, as shown in Figure 2(b), resulting in a more symmetrical data distribution.

Fig. 2(a): The Distribution of Original CBR Values

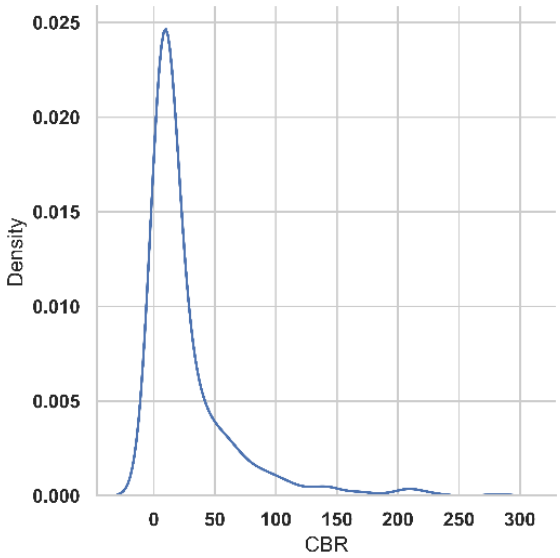
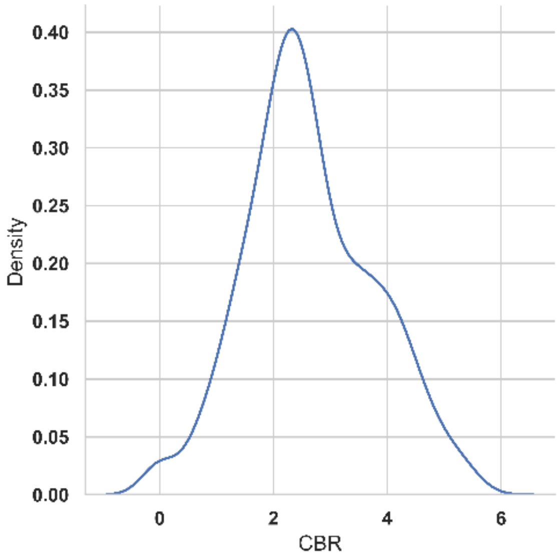


Fig. 2(b): The Distribution of Transformed CBR Values



In addition to the density plots, box plots were used to show the distribution of the data as depicted in Figure 3(a)-(h). Box plots provide a detailed overview of data distribution, illustrating the range, outliers, maximum and minimum values, as well as the median, enhancing the understanding of data spread. In this study, outliers were identified and carefully dropped from the dataset to ensure accuracy and avoid skewing the model results.

Figure 3(a) Liquid Limit Box Plot

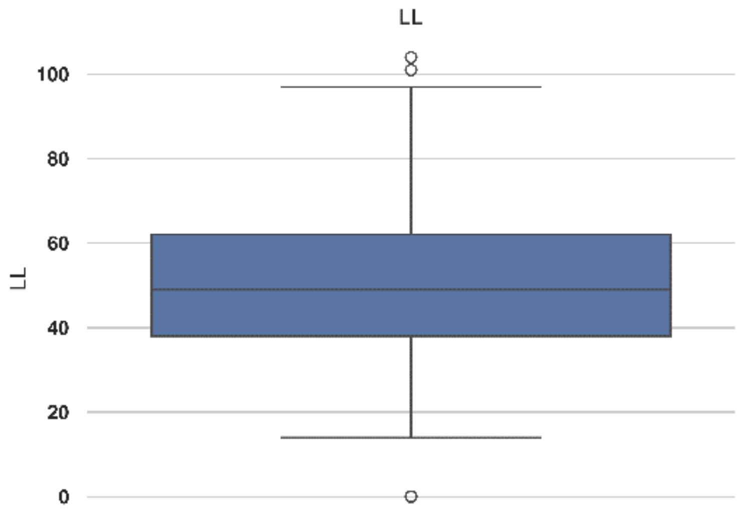


Figure 3(b): Plastic Limit Box Plot

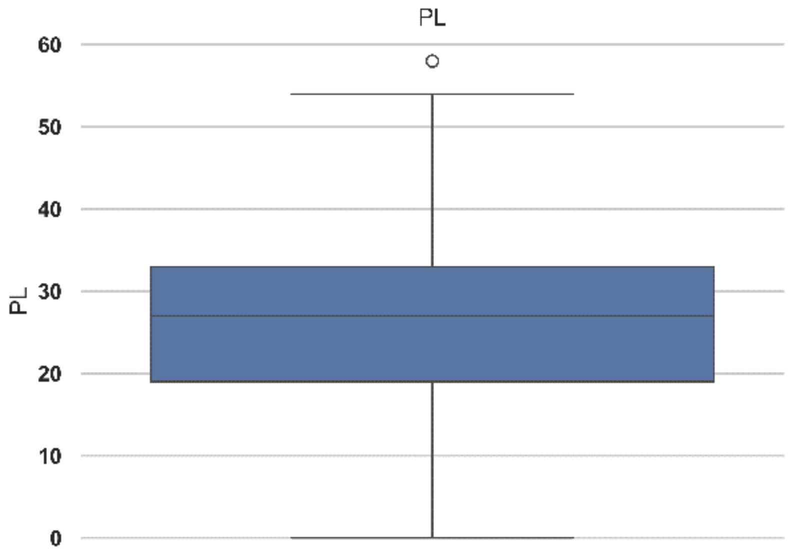


Figure 3(c): Plasticity Index Box Plot

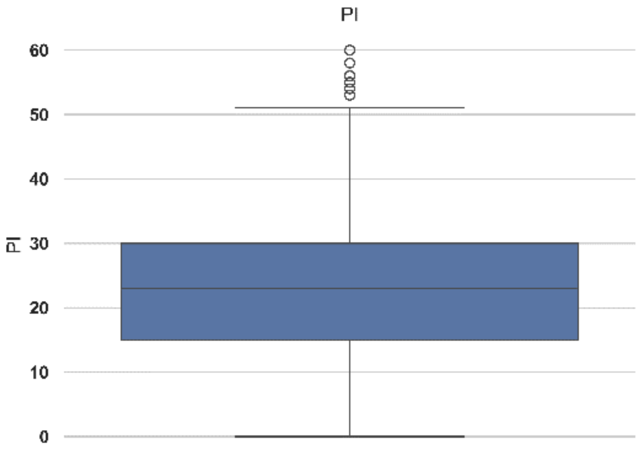


Figure 3(d): Maximum Dry Density Box Plot

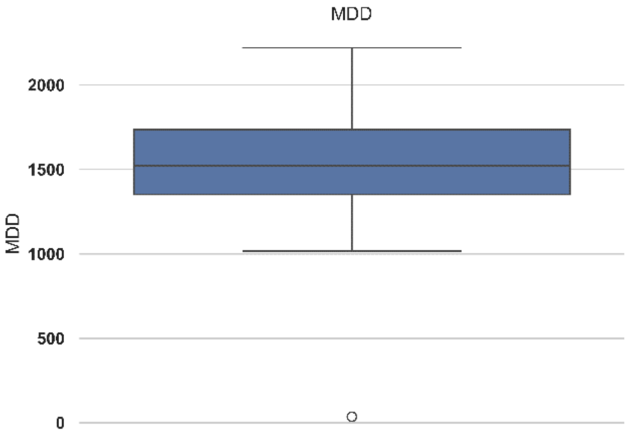


Figure 3(e): Optimum Moisture Content Box Plot

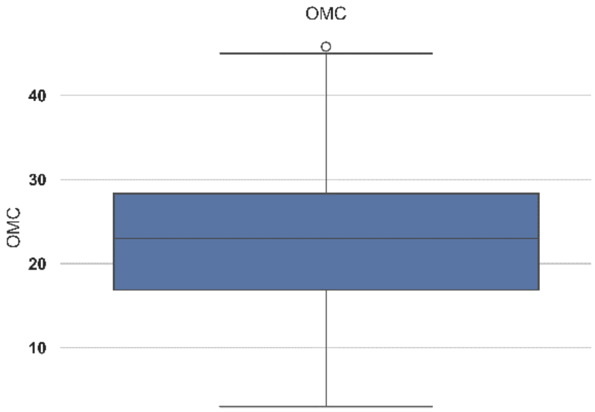


Figure 3(f): The Percentage of Soil Particles Passing through the 0.425mm Diameter Sieve

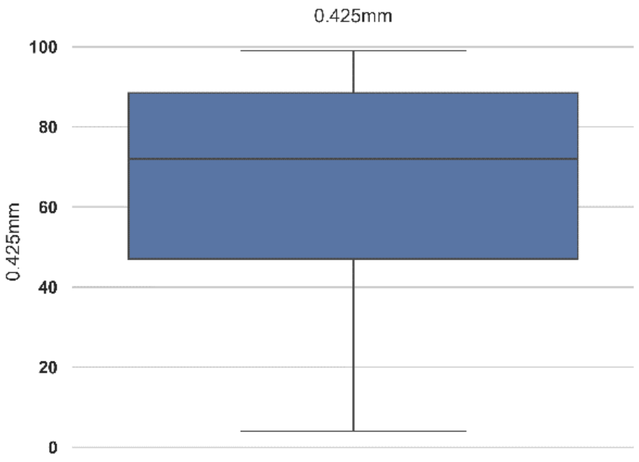


Figure 3(g): The Percentage of Soil Particles Passing through the 0.075mm Diameter Sieve

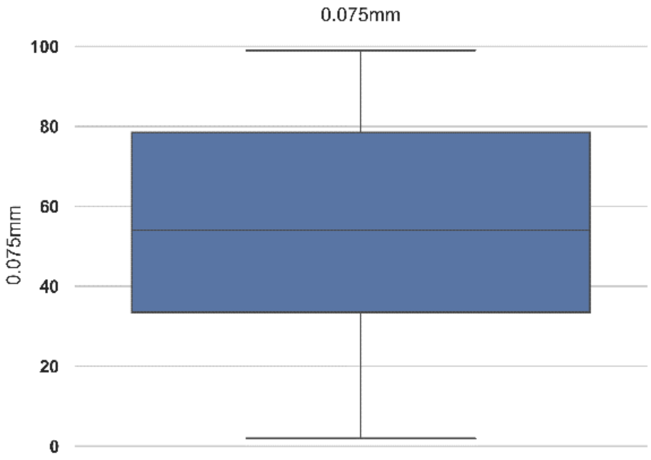


Figure 3(h): California Bearing Ratio Box Plot

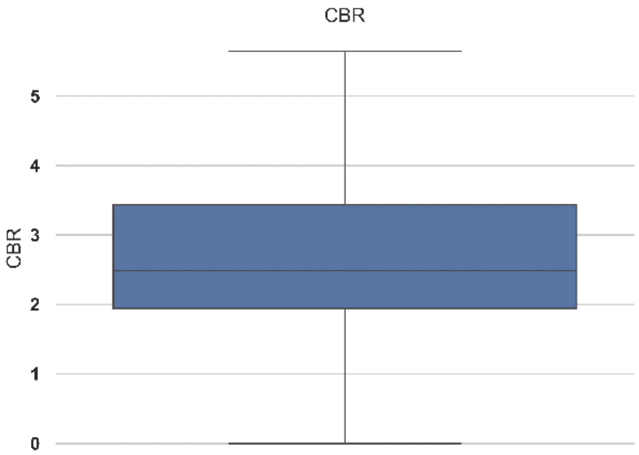
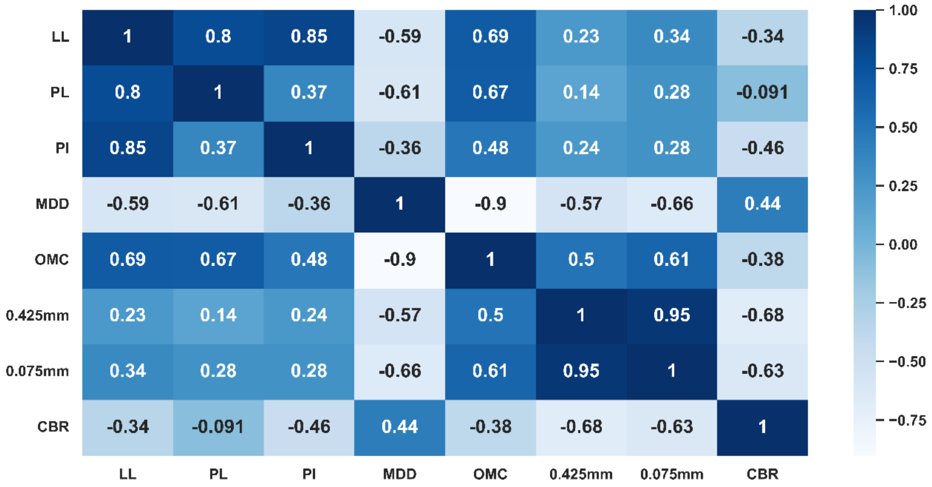


Figure 4 shows the correlation between the different variables in the dataset, highlighting the strength and direction of relationships.

Figure 4: Correlation of Variables



	LL	PL	PI	MDD	OMC	0.425mm	0.075mm	CBR
LL	1	0.8	0.85	-0.59	0.69	0.23	0.34	-0.34
PL	0.8	1	0.37	-0.61	0.67	0.14	0.28	-0.091
PI	0.85	0.37	1	-0.36	0.48	0.24	0.28	-0.46
MDD	-0.59	-0.61	-0.36	1	-0.9	-0.57	-0.66	0.44
OMC	0.69	0.67	0.48	-0.9	1	0.5	0.61	-0.38
0.425mm	0.23	0.14	0.24	-0.57	0.5	1	0.95	-0.68
0.075mm	0.34	0.28	0.28	-0.66	0.61	0.95	1	-0.63
CBR	-0.34	-0.091	-0.46	0.44	-0.38	-0.68	-0.63	1

Predictive Models

Random Forest (RF) Model

The random forest model is a robust and highly versatile machine-learning technique within the ensemble learning family. Introduced by Kwok and Carter in 1990, it builds multiple decision trees during the training process to enhance predictive accuracy. Unlike traditional decision trees, which may suffer from overfitting when dealing with complex datasets, random forest reduces overfitting by averaging the outputs of each tree, thereby improving stability. (Kwok, & Carter, 1990).

The "random" in random forest refers to two essential elements: the random sampling of training data to construct each tree and the random selection of features at each split during the tree-building process, which introduces beneficial variability. These randomization techniques effectively improve the generalization performance and robustness of the RF models, enabling them to handle high-dimensional data more efficiently (Schonlau, & Zou, 2020).

Linear Regression

Linear regression is a commonly used statistical method aimed at modelling the relationship between a dependent variable and one or more independent variables, assuming that this relationship is linear (Huang, 2022). It is one of the foundational tools in data analysis due to its simplicity and interpretability.

Linear regression can be categorized into two types, distinguished by the number of independent variables involved. These are simple and multiple linear regression. Simple linear regression models the relationship between exactly two variables, one dependent and one independent, while multiple linear regression models the relationship between one dependent variable and multiple independent variables simultaneously. The main aim of this technique is to find the best-fitting straight line or hyperplane in higher dimensions, which minimizes the sum of squared residuals and provides the most accurate predictions (Su et al., 2012).

Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a highly efficient and powerful ML model that belongs to the ensemble learning family, similar to the random forest model. It excels in handling both regression and classification tasks. It works by combining simple models, like decision trees, sequentially to improve predictive accuracy with each iteration. Unlike traditional gradient boosting methods, XGBoost uses a more sophisticated and efficient optimization algorithm that minimizes errors through advanced techniques such as regularization. This technique enhances overall model performance by penalizing complexity, therefore effectively reducing overfitting, which is common in complex datasets.

This model has several innovative features, such as parallel and distributed computing capabilities, tree pruning, early stopping, and support for custom loss functions. These advanced features make the model highly effective for tackling a wide range of regression problems, including those with large datasets and high dimensionality. Its remarkable speed and accuracy have made it a popular choice for various modelling requirements, particularly in competitive data science. Additionally, it offers interpretable results, which makes it easy for users to understand feature importance and model behaviour, providing valuable insights into the data (Friedman et al., 2002).

Evaluation Metrics

The performance and accuracy of the three models were evaluated using four reliable metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R²).

MSE calculates the average of the squared differences between predicted and actual values, while RMSE is the square root of MSE. By expressing errors in the same units as the target variable, RMSE is easier to interpret.

$$MSE = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \qquad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad (2)$$

MAE represents the mean of the absolute differences between predicted and actual values.

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

The R² value represents the proportion of variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, where 1 signifies perfect predictive accuracy, while 0 indicates no predictive capability (Cheng et al., 2014).

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

RESULTS

The models were trained on 80% of the dataset and then evaluated on the remaining 20% of the data to assess their predictive performance. The accuracy and effectiveness of these models were carefully evaluated using four key metrics: MAE, MSE, RMSE, and R². The results of this evaluation are presented in Table 3 and Figures 5(a)-(d), providing a clear visual and tabular representation of the model performance.

Table 3: Results of the Study

	MAE	MSE	RMSE	R²
Random Forest	9.21	297.58	17.25	0.74
Linear Regression	12.16	712.89	26.7	0.36
XGBoost	7.98	194.39	13.94	0.82

Figure 5 (a): Mean Absolute Error (MAE)

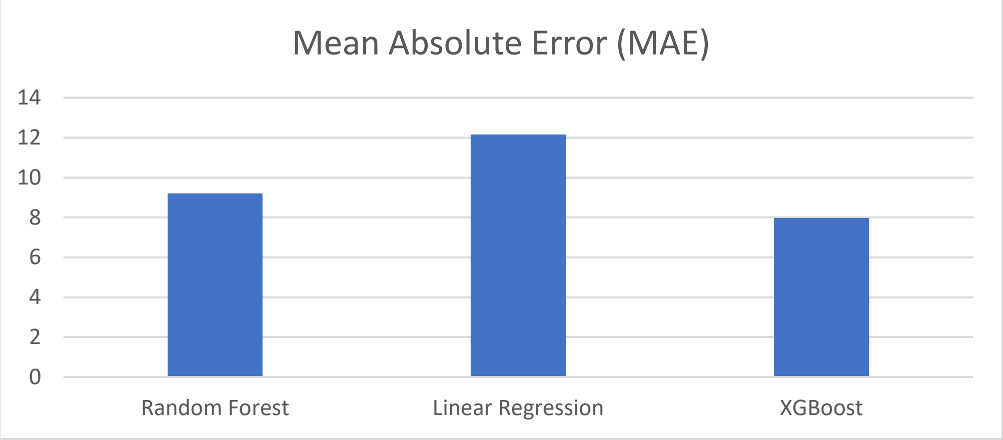


Figure 5 (b): Mean Squared Error (MSE)

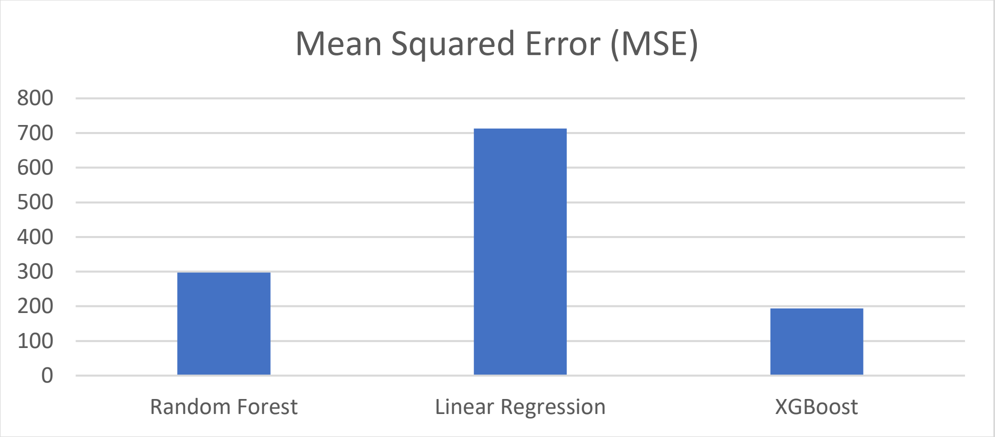
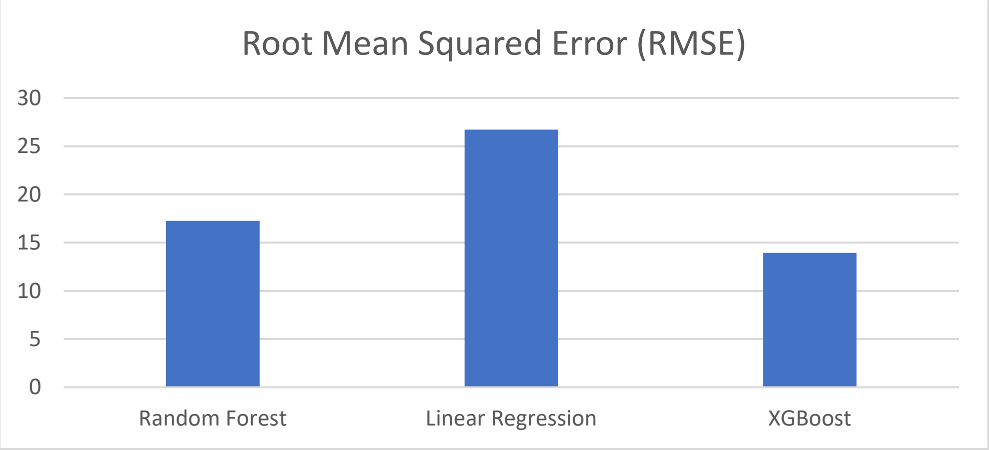
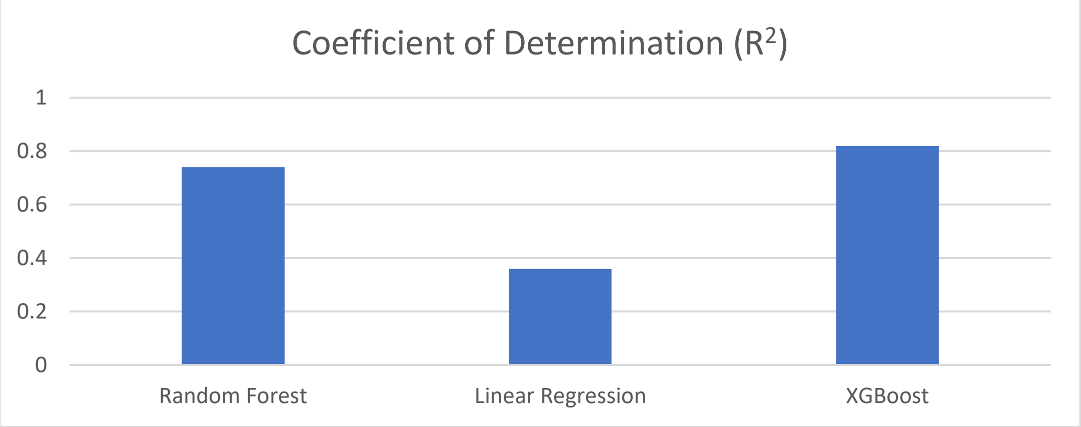


Figure 5 (c): Root Mean Squared Error (RMSE)



Article DOI: https://doi.org/10.37284/eaje.8.1.2651

Figure 5 (d): Coefficient of Determination (R²)



From these results, it becomes evident and undeniable that the XGBoost model stands out as the most accurate and reliable among all the models tested. It achieved the lowest values for MAE, MSE, and RMSE, while also securing the highest R² value, reinforcing its superior predictive accuracy and robustness.

DISCUSSION

It is not always possible to definitively conclude that a single model is universally the most accurate, as the performance and effectiveness of various models can significantly vary depending on factors such as the specific dataset being used and the input variables selected in different studies or applications.

However, in this particular study, the XGBoost model emerged as the most accurate model among those tested, as it demonstrated the lowest MAE, MSE, and RMSE, with respective values of 7.98, 194.39, and 13.94. Additionally, it achieved the highest coefficient of determination value of 0.82, further highlighting its superior predictive capabilities. As a result, the XGBoost model was ultimately selected and utilized to predict the CBR values. The outcomes of the predicted values are visually presented in Figure 6(a)-(b), showcasing the model's performance.

Figure 6 (a): Training Data Scatter Plot

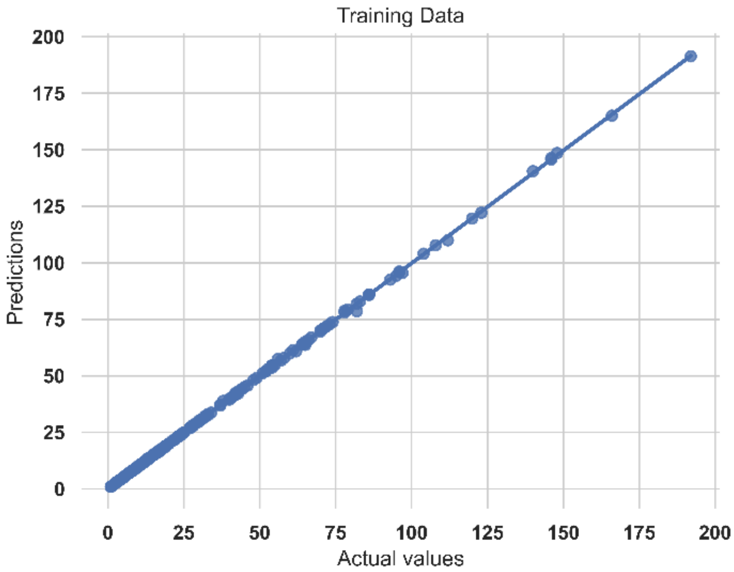
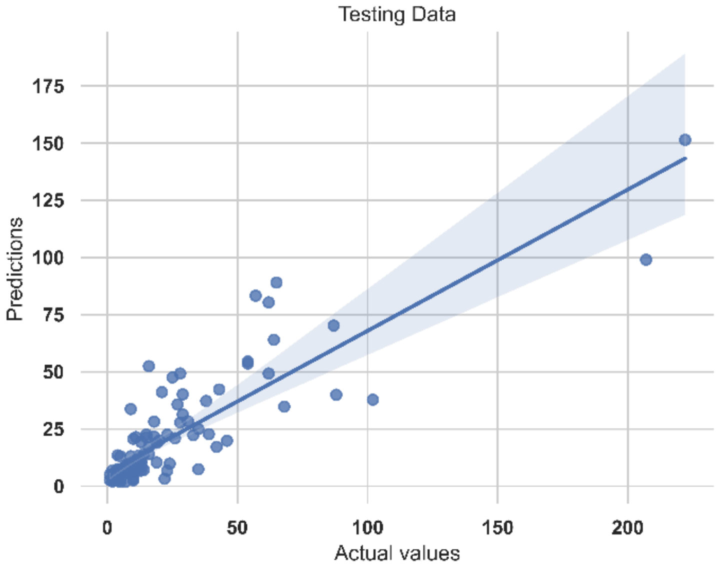


Figure 6 (b): Testing Data Scatter Plot



CONCLUSION

This study aimed to comprehensively address the inherently time-consuming nature of the California Bearing Ratio (CBR) test by leveraging modern machine learning techniques to accurately predict CBR values. To achieve this objective, three distinct models—random forest, linear regression, and extreme gradient boosting (XGBoost)—were systematically trained on 80% of the dataset and subsequently tested on the remaining 20%. Among these models, the XGBoost model stood out as the most accurate and reliable, achieving the lowest mean absolute error (MAE) of 7.98, mean squared error (MSE) of 194.39, and root mean squared error (RMSE) of 13.94. Additionally, it demonstrated the highest coefficient of determination (R^2) of 0.82, stressing its superior predictive performance.

Looking forward, future studies should explore the practical deployment of this XGBoost model within a decision support system (DSS) framework to assess its tangible impact on real-world applications. Furthermore, continuous and consistent data collection on the CBR dataset is highly recommended, as it would facilitate training with more advanced and sophisticated models, including cutting-edge approaches such as artificial neural networks (ANN). This ongoing effort to improve and refine the predictive modelling process has the potential to significantly enhance the efficiency and accuracy of CBR predictions, thereby addressing key challenges in soil strength evaluation and construction project planning.

REFERENCES

ASTM D1883-16. (2016). Standard Test Method for California Bearing Ratio (CBR) of Laboratory-Compacted Soils. *Astm International*.

Çelik, Ö. (2018). A Research on Machine Learning Methods and Its Applications. *Journal of Educational Technology and Online Learning*. https://doi.org/10.31681/jetol.457046

Cheng, C. L., Shalabh., & Garg, G. (2014). Coefficient of determination for multiple measurement error models. *Journal of Multivariate Analysis*. https://doi.org/10.1016/j.jmva.2014.01.006

Firican, G. (2022). *The History of Machine Learning*. Lights On Data.

Friedman, J., Hastie, T., & Tibshirani, R. (2002). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*. https://doi.org/10.1214/aos/1016218223

Huang, S. (2022). Linear regression analysis. In *International Encyclopedia of Education: Fourth Edition*. https://doi.org/10.1016/B978-0-12-818630-5.10067-3

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*. https://doi.org/10.1007/s12525-021-00475-2

Keith D. Foote. (2019). *A Brief History of Machine Learning*. Https://Www.Dataversity.Net/a-Brief-History-of-Machine-Learning/.

Kwok, S. W., & Carter, C. (1990). Multiple decision trees. In *Machine Intelligence and Pattern Recognition*. https://doi.org/10.1016/B978-0-444-88650-7.50030-5

Nguyen, B. T., & Mohajerani, A. (2015). Prediction of California Bearing Ratio from Physical Properties of Fine-Grained Soils. *International Journal of Civil, Structural, Construction and Architectural Engineering*, *9*(2), 136– 141. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.674.2360&rep=rep1&type=pdf

Sasakawa, T., Hu, J., & Hirasawa, K. (2008). A brainlike learning system with supervised, unsupervised, and reinforcement Learning. *Electrical Engineering in Japan (English Translation of Denki Gakkai Ronbunshi)*. https://doi.org/10.1002/eej.20600

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*. https://doi.org/10.1177/1536867X20909688

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.1198

Sylvia, M. L., & Murphy, S. (2023). Exploratory Data Analysis. In *Clinical Analytics and Data Management for the DNP, Third Edition*. https://doi.org/10.1891/9780826163240.0014

Yoder, E. J., M. W. Witczak, Witczak, M. W., & M. W. Witczak. (1975). Principles of Pavement Design, Second Edition. *Principles of Pavement Design*.