International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Prediction of California Bearing Ratio Using Machine Learning for Stabilized Soil Subgrade

K. Harsha Vardhan [a], K. Sharmila Rani [b], Ch. Veera Naga Babu [c], G. Satyanarayana [d]

[a,b,c,d] UG Student, GMR Institute of Technology, Rajam and 53212, India

ABSTRACT

The California Bearing Ratio (CBR) is an important engineering measure in the design and evaluation of pavement and foundation systems. To streamline pavement construction and maintenance, ensure traffic safety, and reduce infrastructure costs, an accurate prediction of the CBR value is critical. This abstract provides an overview of an investigation into the use of machine learning to predict CBR values. Real-time or large-scale CBR estimation is difficult because CBR assessment typically requires expensive and time-consuming laboratory experiments. In this study, machine learning methods are used to build predictive models that can estimate CBR values based on readily available geotechnical and soil data. The dataset used for model training consists of a variety of soil samples obtained from different locations in California, covering different soil types and environmental variables. Machine learning models are trained and validated using soil composition, moisture content, particle size distribution, and compaction characteristics. To create CBR prediction models, several machine learning methods are used, including but not limited to Random Forest, Support Vector Machine, and Gradient Boosting. Metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) are used to measure these models' performance to judge their precision and generalizability.

Keywords: California bearing Ratio, *Random Forest, Regression Analysis, Decision Trees, Extreme Gradient Boosting.*

1. Introduction

The CBR value is a crucial parameter in geotechnical engineering used to evaluate the strength and bearing capacity of subsoils. Using machine learning, CBR values can be predicted based on soil properties and environmental factors. This can help engineers assess the suitability of a site for construction projects, evaluate risks, estimate construction costs, and conduct environmental impact assessments. CBR predictions can also be used for quality control during construction and for maintenance planning of existing infrastructure. In a study by researchers at the Indian Institute of Technology Madras, machine learning was used to predict CBR values of fine-grained soils based on their index properties (**Semachew Molla Kassa , 2023)**. The model was able to predict the CBR value with 90% accuracy. Another study by researchers at the College of California, Davis, used machine learning to predict the CBR values of sands based on their particle size distribution and compaction properties. The model was able to predict CBR values with 85% accuracy (Mohammadreza Ghajarinejad, 2023).

Civil engineers, especially those involved in pavement construction, rely on the California Bearing Ratio (CBR) test as a dependable method to assess the strength of the subgrade in comparison to crushed stone aggregates. The CBR test is a key geotechnical test used to evaluate the bearing capacity of soil subgrades for road construction. It measures the resistance of the soil to the penetration of pressure and provides a CBR value that indicates how well the soil can support loads compared to a reference material. This value guides engineers in designing stable and durable pavements by evaluating soil strength, deformation potential, and the need for stabilization methods. It essentially measures the relative strength of soil compared to standard gravel. Traditionally, CBR testing was known to be time-consuming, costly, and labor-intensive, which often affected the efficiency of pavement designs. Predicting CBR values based on soil index properties such as grain size, Atterberg limits, moisture content, and compaction characteristics has proven to be a practical alternative. Granular soils such as sand and gravel are preferred for road construction because of their high CBR values, which indicate good bearing capacity and drainage properties. Cohesive soils such as clay and silt typically have lower CBR values and may need to be stabilized or modified to be suitable for road construction. CBR testing is versatile and can assess the strength of mixed soils commonly found in natural conditions. Engineers use fill materials specifically designed for road construction and often subject them to CBR testing to meet project requirements. CBR testing is essential for evaluating the strength of the subgrade beneath road layers and determining if it needs to be improved or stabilized. Recycled materials such as crushed concrete or asphalt can be used in road construction and are evaluated using CBR testing to ensure they meet standards.

Machine learning is a subfield of artificial intelligence that allows computers to learn from data and improve their performance over time. It involves developing algorithms and models that allow systems to recognize patterns, make predictions, and take actions without being explicitly programmed. By analyzing large amounts of data, machine learning algorithms can gain insights, make informed decisions, and automate complex tasks. This technology is finding applications in fields ranging from healthcare and finance to image recognition and natural language processing, revolutionizing the way we solve problems and harnessing the potential of data-driven intelligence.

1.1 Applications of prediction of California bearing ratio:

- Road pavement design and construction are two of the main areas where CBR prediction is used.
- Engineers use CBR values to select pavement thickness and materials for roads to handle expected traffic loads.
- CBR prediction is essential for assessing the condition of existing road pavements.
- It helps in planning and prioritizing rehabilitation and maintenance activities to extend the life of roads.
- To design the foundations for buildings, bridges, and other structures, engineers employ CBR values.
- In order to guarantee the stability and safety of these structures, accurate CBR predictions are essential.
- CBR values are used by engineers for designing embankments for levees, dams, and retaining walls, among other uses.
- Predictions that are accurate aid in avoiding failure and instability. CBR prediction assists in determining the need for soil improvement techniques, such as soil stabilization or reinforcement, in construction projects.
- CBR prediction is a valuable tool for research in soil mechanics and geotechnical engineering.
- It is also used for educational purposes to teach students about soil properties and their practical applications.
- Predicting the California Bearing Ratio (CBR) using machine learning for stabilized soil subgrade provides a modern approach to evaluating soil strength and bearing capacity. By using advanced algorithms, this method improves the accuracy of estimating CBR values based on various soil properties and stabilization techniques. This innovative technique aims to streamline pavement design and construction by providing engineers with reliable predictions for optimized subgrade performance.
- The use of machine learning to predict the California Bearing Ratio (CBR) in stabilized subgrade represents an innovative method in geotechnical engineering. By using data-driven algorithms, this approach improves the accuracy of CBR estimates, particularly in the context of soil stabilization techniques. To optimize pavement design and construction through more reliable and efficient subgrade evaluation.

1.2 Algorithm:

An algorithm is a precise set of step-by-step instructions or a well-defined procedure for performing a specific task or solving a particular problem. It's a computational or problem-solving strategy that outlines the logical sequence of operations to be followed to achieve a desired result. Algorithms are used in various fields of computer science, mathematics, and engineering, and they serve as the foundation for computer programs and automated processes. In most cases, algorithms start with some input data. This input could take the form of text, photos, numbers, or any other kind of data. Algorithms describe a set of precise actions or operations that methodically alter the incoming data. Programming languages, flowcharts, and pseudocode are frequently used to describe these stages. An algorithm generates an output or result after performing the steps required, which is the answer to the question or the desired result. Algorithms are deterministic, which means that they always result in the same output for a given set of inputs. Their behavior is not arbitrary or ambiguous. Algorithms must be finite, which means they must have a distinct finish. They do not run endlessly but eventually come to an end. For an algorithm to be effective, it must solve the problem correctly and effectively, and it must do it with a minimal amount of time and resources.

1.3 Ensemble:

An ensemble is a machine learning technique that involves combining the predictions of multiple individual models to create a single, more powerful predictive model. Aggregating the outputs of several models enhances overall prediction accuracy and generalization by leveraging the strengths of one model to mitigate the weaknesses of others. Combining the predictions of various separate models, also known as base models or weak learners, into an ensemble technique is a potent machine learning strategy that results in a final prediction that is more reliable and accurate. In order to increase overall prediction performance and lower the risk of overfitting, ensemble approaches aim to take advantage of the variety of these base models. When doing several machine learning tasks, like as classification, regression, and even CBR prediction in geotechnical engineering, ensemble methods are frequently utilized.

Some standard ensemble algorithms include Bagging, Random Forests, Boosting, Gradient Boost (e.g., Extreme Gradient Boost (XG Boost), Light GBM, AdaBoost), and Stacking.

Random Forest Regression is an ensemble method for predicting continuous values. It creates a collection of decision trees, each trained on random data subsets and feature subsets. During prediction, it combines these tree outputs, often by averaging, to build a strong and reliable regression model. This approach is adept at handling complex relationships, preventing overfitting, and assessing feature importance. It's versatile, robust against outliers, and applicable in various regression scenarios, such as finance, sales, and scientific research. Multiple Linear Regression is a statistical method that models the relationship between a dependent variable and multiple independent variables through a linear equation. It aims to find the best-fitting equation to explain variations in the dependent variable, helping assess the strength and direction of relationships for prediction and hypothesis testing. However, it requires validating assumptions like linearity, independence, and normality of residuals. This technique is widely used in fields like economics, social sciences, and data analysis to uncover complex associations among multiple factors. XG Boost is a potent ensemble algorithm for classification and regression, creating accurate models by combining sequential decision trees. It employs gradient boosting to iteratively reduce prediction errors, handling missing data efficiently, applying regularization for overfitting prevention, and capturing complex data relationships. XGBoost's exceptional performance, speed, and versatility have made it a go-to choice across industries like finance, healthcare, and recommendations for robust predictive modeling. Gradient Boosting is a powerful machine learning technique that creates robust predictive models by repeatedly mixing weak learners, frequently decision trees. By instructing each new learner on the residuals of the prior model, it gradually corrects errors up until a predetermined number of models are produced or a performance threshold is reached. When dealing with complex data relationships, gradient boosting excels at producing precise predictions for both regression and classification problems. It establishes itself as an important machine learning tool by finding extensive applications in fields including ranking, suggestions, and predictive modelling. A decision tree is a fundamental machine learning tool for classification and regression tasks, using a hierarchical structure where nodes represent feature-based decisions, and leaves signify predictions. By recursively dividing data based on informative features to maximize information gain or reduce impurity, decision trees offer interpretability, aiding in feature importance and decision understanding. Nonetheless, they may overfit when deep, necessitating methods like pruning for mitigation. Their simplicity and adaptability make decision trees valuable in fields like medicine, finance, and recommendations.

2. Literature Review

S NO	AUTHOR	TITLE OF PAPER	YEAR	METHODOLOGY ADOPTED	OBSERVATION
1	Zohib Shahzad Janjua et al.,	Correlation of CBR with index properties of soil	2016	In the Punjab city of Gharua n, 11 soil samples were gathered within a 50 km radius. Lab experiments adhered to IS code standards, creating connections between experime ntal and anticipated data. Mostly well-graded sand with silt made up the samples.	Soil CBR is influenced by factors including MDD, OMC, LL, unconfined compressive strength, particle percentage, etc. Focusing on SW-SM soil (well-graded sand with silt). Lab-experimented CBR values are used for regression analysis, deriving an equation linking CBR to soil characteristics.
2	Patel, Rashmi S et al.,	CBR Predicted by Index Properties for Alluvial Soils	2010	Laboratory tests analyzed with Excel and SPSS to improve the relationship between California Bearing Ratio, Linear Regression, and Soil Index Properties. Graphs show CBR's sensitivity to soil conditions, rising with plasticity index, max dry density, and optimal moisture content. Unsoaked CBR/soaked CBR ratio is 0.5 in comparison.	Experimental outcomes are being compared to computed results and are being statistically analyzed using SPSS and programming. Within limits, equations for unsoaked and soaked CBR are being developed based on soil traits. This, along with practical insights, is estimating CBR values at 100m intervals. Correlation is being validated by CBR tests showing consistent ranges.
3	Ahmad Taha Abdulsadda et al.,	Predicting CBR Value from Index Properties of Soils using Expert System	2017	They employ the multilayer perceptron (MLP) architecture for our neural network. This structure, comprising input, hidden, and output layers, is commonly utilized for nonlinear classification and prediction tasks.	Presenting a novel expert system (Multilayer perceptron MLP neural network) that acts as a computer decision maker and predicts the exact CBR value based on data.
4	Khalid R. Mahmood Aljanabi et al.,	Using Artificial Neural Networks to predict the Unconfined Compressive Strength of Clayey	2023	Executed ASTM-standard lab tests (LL, PL, SL, UCS) and regular checkups. In-depth study and modeling of ANN techniques: fundamentals, components, applications,	Based on the experimental results: Additives like sandstone powder, iron ore, and shale rock powder are increasing the soil's unconfined compressive

		Soils Stabilized by Various Stabilization Agents		perceptrons, layers, and algorithms.	strength (UCS). However, the remaining additives are decreasing the UCS as the amount of additive grows. Neural networks can capture complicated relationships between elements and allow you to study the impact of each input variable.
5	A. Bharath et al.,	Influence and correlation of maximum dry density on soaked and unsoaked CBR of soil	2021	In accordance with code requirements, determine the different index properties of soil samples, such as liquid limit, plastic limit, specific gravity, standard proctor compaction, and modified proctor compaction.	Results are showing that soil parameters are impacting the CBR value of soil. Correlations with good R^2 values are being constructed between the California bearing ratio and maximum dry density, allowing us to anticipate CBR values.
6	Ravichandra A H et al.,	Prediction of CBR by using index properties of soil	2019	Regression analysis models relationships between variables. Dependent variable is predicted; independent variable forecasts it. Simple Linear Regression (SLRA) connects basic soil properties to soaking CBR value.	Presenting a novel expert system (Multilayer perceptron MLP neural network) that acts as a computer decision maker and predicts the exact CBR value based on data.
7	S.H. Vamsi Krishna et al.,	Prediction of UCS and CBR of a stabilized Black-cotton soil using artificial intelligence approach: ANN	2023	Utilizing soil stabilization techniques like fly ash, lime, and cement to enhance black-cotton soil's structure and minimize expansion and contraction. Evaluating model performance with statistical measures: R-squared (R2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Relative RMSE (RRMSE).	ANN performs better than traditional methods in predicting CBR and UCS for stabilizing black-cotton soil. Additives like fly ash, lime, and cement are reducing the soil's expansibility and contractility. The precise ANN model is forecasting CBR and UCS, considering pertinent inputs for stabilizing the soil. Engineers can manage expansive-soil challenges in road and civil projects using ANN models.
8	Muthu Lakshmi S et al.,	Predicting soaked CBR of SC subgrade from dry density for light and heavy compaction	2020	Following Indian Standards (ISCS), this study conducted standard tests and Standard/Modified Proctor Compaction tests to determine MDD and OMC using light and heavy compaction. Soaked CBR strength was assessed at different compaction levels (97%, 94%, 91%, and 88%) with both light and heavy compaction. Empirical correlations, established via Statistical Linear Regression Analysis (SLRA), predicted soaked CBR values based on	The dry density (DD), which was produced by both mild and severe com paction, is compared to the soaking CBR value of the SC (Clayey Sand) soil. Using empirical correlations, estimate the subgrade soil's wet CBR strength from SC soil. The DD of the subgrade soil can be precisely determined even without performing the CBR test in the lab. In this work, it has also been attempted to investigate the effects of different compaction energies.

				Dry Density (DD), simplifying and reducing the labour-intensive nature of CBR testing.	According to the wet CBR strength of the SC soil.
9	Semachew Molla Kassa et al.,	Use of Machine Learning to Predict California Bearing Ratio of Soils	2023	Study utilized 252 soil samples, 7 predictors (e.g., max dry density, soil classification), 80% training, 20% test data split. Models trained on predictors for CBR estimation. Evaluation metrics (MSE, MAE, RMSE, R2) favoured random forest, showcasing lower errors and higher R2 for robust CBR prediction.	Demonstrating random forest's excellence in accurately predicting soil CBR, outperforming other methods. Metrics such as MSE, MAE, RMSE, and R2 are confirming its lower errors and higher R2 value. The study is underscoring the efficacy of random forest in accurately predicting CBR based on modal elevation.
10	Likhith K M et al.,	Prediction of California Bearing Ratio (CBR) of Soils using AI-based Techniques	2022	Random forest is a robust machine learning algorithm that utilizes multiple decision trees to make predictions. Each tree is trained on a subset of the data, and their predictions are averaged, reducing the risk of overfitting compared to a single tree. Artificial neural networks (ANNs) are another machine learning method composed of interconnected neurons. Each neuron specializes in a specific task, allowing ANNs to model intricate relationships between input and output variables.	RFR model had good predictive performance (R2=0.92, MSE=16.2), while the ANN model performed slightly better in terms of correlation (R=0.95) but had a higher mean squared error (MSE=28). A sensitivity analysis identified Maximum Dry Density (MDD) and the percentage of fines as the most influential soil parameters on CBR, with Plastic Limit (PL) having the least impact.

Results and Discussion

Linear Regression Analysis:

Linear regression predicts outcomes from variable values, ideal for proportional connections. It calculates the best-fit line (a + bx) to minimize prediction differences by determining intercept (a) and slope (b). Used widely for modelling, it hinges on linear assumptions and consistent data. Simplicity and interpretability: Linear regression provides straightforward interpretations of the relationships between input features and CBR values. Fast training and prediction. Assumes a linear relationship between input features and CBR, which may not hold in complex geotechnical scenarios. Limited ability to capture nonlinear patterns.

Multiple Regression Analysis:

Multiple regression is a way to understand how one thing depends on several others. Imagine you're trying to predict something, like how fast a car goes, and you have lots of factors like engine size and weight. Multiple regression helps find the best math equation to explain how all these factors work together to affect the car's speed. It's really useful for making predictions and figuring out which factors matter the most. But, just like with regular regression, we have to follow certain rules to get accurate answers. In this kind of math, we're trying to make a straight line that fits the data points as closely as possible. This line helps us predict the thing we're interested in based on the other factors we have.

Random Forest:

A Random Forest is a machine-learning technique that combines multiple decision trees to enhance prediction accuracy. It's used for classification and regression tasks. In a Random Forest, each tree is trained on a unique data subset, and their combined predictions yield more reliable results. Random Forest Regression is a machine learning method used for regression tasks, and it is frequently abbreviated as Random Forest. It is an ensemble learning technique that harnesses the potential of numerous decision trees to provide more accurate predictions. Due to its durability and capacity to handle a wide range of data sources and challenging regression issues, Random Forest is a flexible and frequently used method in data science and machine learning. Compared to individual decision trees, it is resilient and less prone to overfitting. Numerous data kinds, such as numerical and categorical features, can be handled by it. It offers feature importance ratings that can be used to pick features and gauge the relative weights of several features in a prediction.

Decision Tree:

Random Forest and gradient-boosting models can be hard to understand when lots of decision trees are combined. Decision trees are like flowcharts that help us make choices or predictions in machine learning. They break down a big decision into smaller steps. We use them a lot for sorting things into categories or predicting numbers. We build decision trees by splitting the data into smaller groups at each step, trying to make these groups as similar as possible. This helps us make good decisions or predictions. Decision trees work well with different types of data and can handle missing information. But they can get too complicated and make mistakes if we're not careful. We can fix this by using methods like Random Forests, which are like a group of decision trees working together. Decision trees are important in machine learning and are used in many other fancy models. They can find complex patterns in data but can be tricky if not used correctly.

Extreme Gradient Boosting (XG Boost):

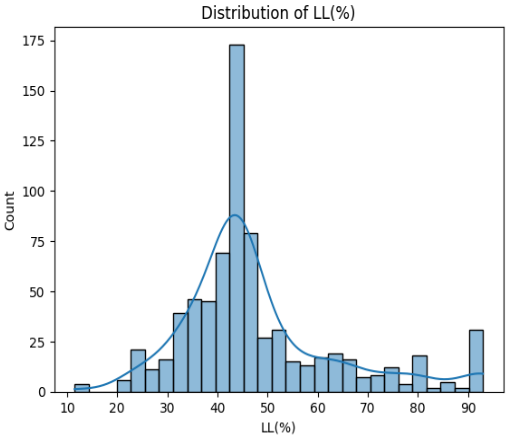
A machine-learning technique called XG Boost can be applied to a wide range of issues. It is simple, straightforward, and effective. XG Boost is a wonderful option if you're seeking a machine-learning algorithm that can produce successful outcomes. Both classification and regression issues can be solved using XG Boost. The objective of regression problems is to predict a continuous value. The objective of classification issues is to forecast a categorical value. XG Boost is renowned for its outstanding prediction performance. It frequently performs better than other machine learning algorithms in a variety of jobs. Large datasets and high-dimensional feature spaces can be handled, and it is computationally effective. Because of its regularization methods, XG Boost is resistant to outliers and noisy data. It works well with both structured and unstructured data and can be applied to both classification and regression problems. XG Boost offers feature importance, which enables users to comprehend the significance of each feature in the model's decisions even though it is less interpretable than linear models. To attain the best performance, XG Boost contains various hyperparameters that need to be carefully tuned. This may take a lot of time and expertise. Training XG Boost models may require a lot of memory and processing power for very large datasets.

Gradient Boosting:

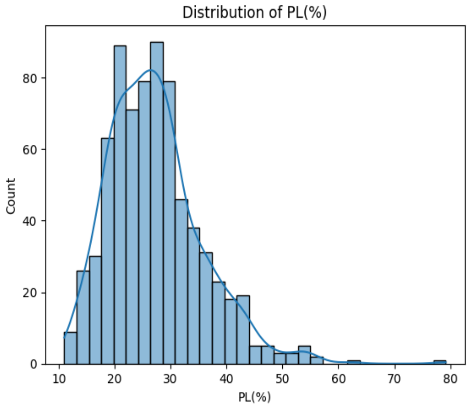
Gradient Boosting is an effective ensemble machine-learning method used for both classification and regression tasks. It is renowned for its capacity to create incredibly precise predictive models by progressively training several weak learners (often decision trees) and integrating their predictions in a weighted way. Gradient boosting algorithms, including Gradient Boosting Machines (GBM) and XG Boost, have gained popularity because of their cutting-edge performance in a variety of applications. Gradient Boosting has frequently taken first place in machine learning competitions due to its outstanding prediction accuracy. It is appropriate for A range of problem areas since it can handle a wide range of data types and complexities. Because the optimization process focuses on reducing mistakes, gradient boosting is resistant to outliers and noisy data. Users can comprehend how each feature contributes to the prediction of the model by using the measure of feature importance it offers. When using a lot of base learners or deep trees, gradient-boosting models might get complicated, which could result in longer training durations and overfitting.

	Measurement	N	Minimum	Maximum	Mean	Std. deviation	variance
LL	%	30	20	84	49	17	290
PL	%	30	11	79.2	27.2	8.4	71.88
OMC	%	30	1.64	17.2	22.6	62.7	3838
MDD	g/cc	30	1.24	23.4	5.16	6.57	43.22
CBR	%	30	1	15	11.9	14.06	197.925

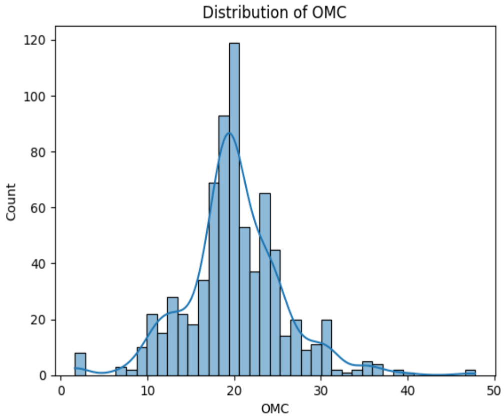
Table. 1: statistical information of dependent variables and independent variables

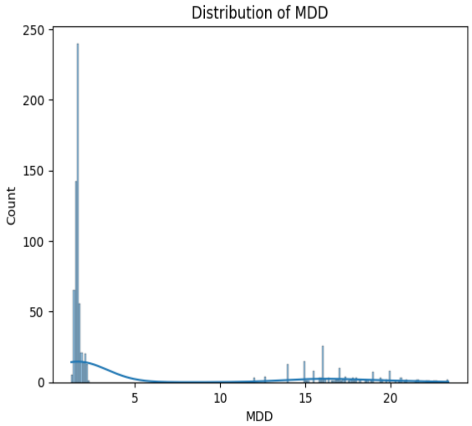


Graph 1: Distribution of Liquid Limit



Graph 2: Distribution of Plastic Limit





Graph 3: Distribution of Optimum Moisture Content

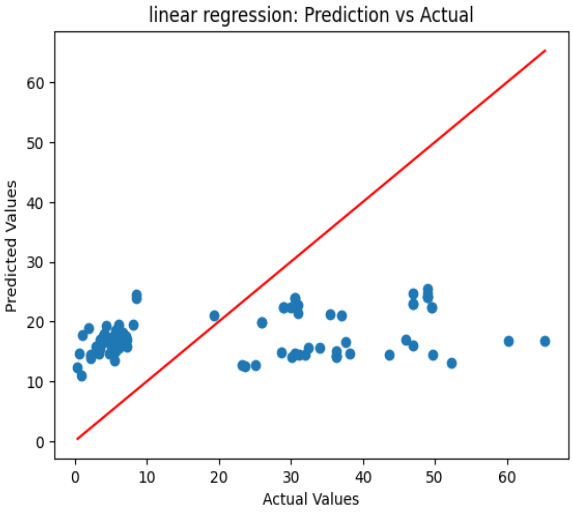
Graph 4: Distribution of Maximum Dry Density

Linear Regression:

Datasets	R^2	Root Mean Squared Error	Mean Squared Error	Median Absolute Error	Variance
Training	0.07	17.03	290.1	12.19	0.051
Testing	0.05	17.32	300.2	12.35	0.095
All	0.07		300.25	12.36	0.1

Table. 2: Statistical evaluation of the performance of Linear Regression

In a linear regression model, the R^2 indicates that approximately 7% of the variance in the dependent variable is explained by the independent variable. The root mean squared error (RMSE) for the model is 17.32, suggesting the average prediction error of approximately 17.32 units on the testing dataset.



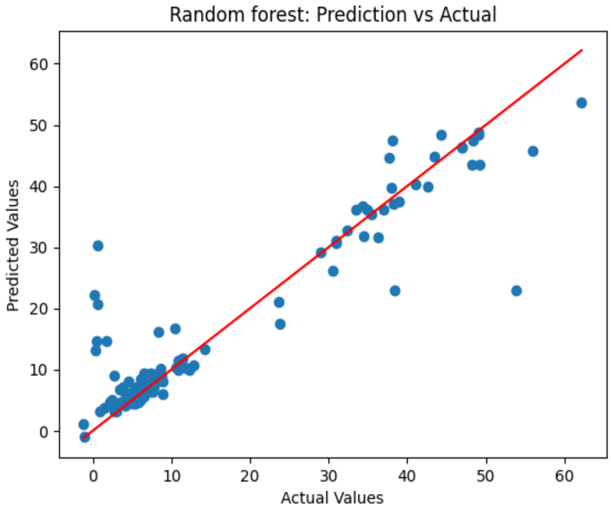
Graph 5: Linear Regression

Random Forest:

Datasets	R^2	Root Mean Squared Error	Mean Squared Error	Median Absolute Error	Variance
Training	0.98	1.35	1.82	0.43	0.98
Testing	0.90	4.80	23.07	0.85	0.90
All	0.9		23.07	0.85	0.9

Table. 3: Statistical evaluation of the performance of Random Forest

In the Random Forest model, an impressive R2 value of 0.98 on the training dataset demonstrates a high proportion of variance explained, with a low RMSE of 1.35 indicating accurate predictions. On the testing dataset, an R^2 of 0.90 suggests robust generalization, despite a slightly higher RMSE of 4.80, indicating moderate prediction errors.



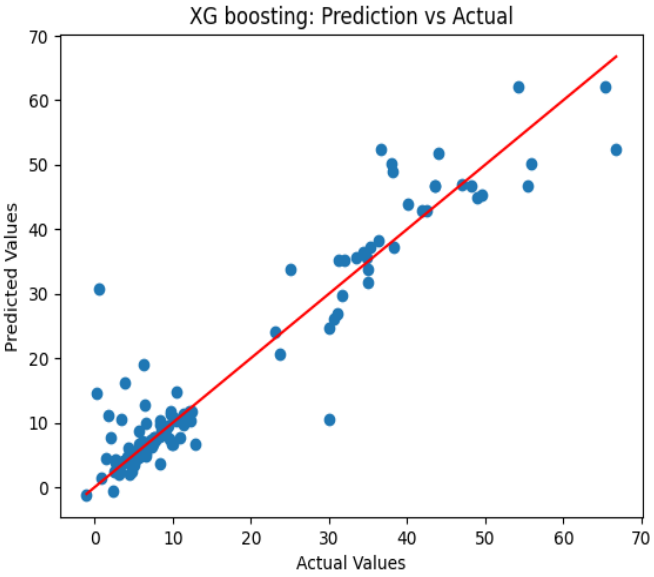
Graph 6: Random Forest

Extreme Gradient Boosting:

Datasets	R^2	Root Mean Squared Error	Mean Squared Error	Median Absolute Error	Variance
Training	0.98	2.00	4.01	0.35	0.98
Testing	0.87	5.12	26.28	0.64	0.87
All	0.88		26.29	0.65	0.88

Table. 4: Statistical evaluation of the performance of Extreme Gradient Boosting

In Extreme Gradient Boosting, the model exhibits strong training performance, with an R^2 of 0.98, indicating a high proportion of variance explained, and a relatively low RMSE of 2.00, signifying accurate predictions. However, on the testing dataset, it shows a slightly reduced R^2 of 0.87 and a higher RMSE of 5.12, suggesting good generalization with a somewhat increased prediction error.



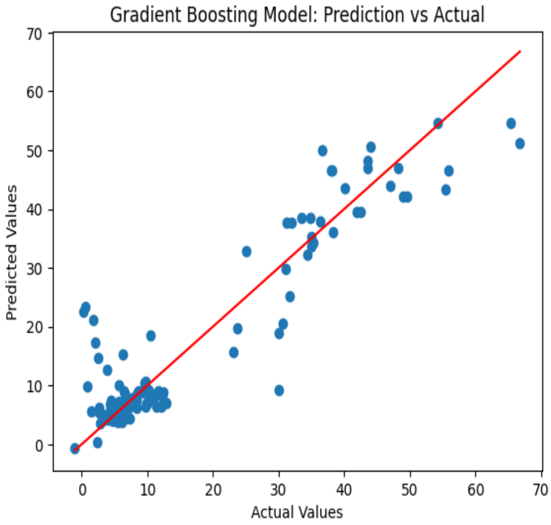
Graph 7: Extreme Gradient Boosting

Gradient Boosting:

Datasets	R^2	Root Mean Squared Error	Mean Squared Error	Median Absolute Error	Variance
Training	0.96	2.66	7.08	1.05	0.96
Testing	0.87	5.46	29.90	1.26	0.87
All	0.87		29.9	1.26	0.87

Table. 5: Statistical evaluation of the performance of Gradient Boosting

In Gradient Boosting, the model demonstrates strong training performance, achieving an R^2 of 0.96, indicating a high degree of explained variance, and a relatively low RMSE of 2.66, signifying accurate predictions. However, on the testing dataset, it exhibits a slightly reduced R^2 of 0.87 and a higher RMSE of 5.46, implying good generalization but with a somewhat increased prediction error.



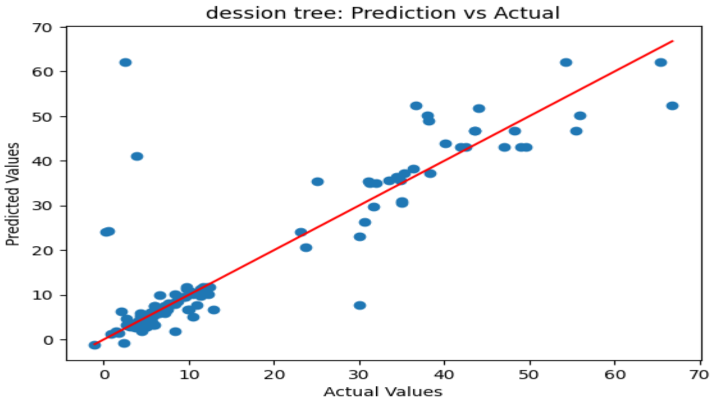
Graph 8: Gradient Boosting Model

Decision Tree:

Datasets	R^2	Root Mean Squared Error	Mean Squared Error	Median Absolute Error	Variance
Training	0.99	1.35	1.82	0.0	0.99
Testing	0.77	7.37	54.34	0.533	0.77
All	0.77		54.35	0.53	0.77

Table. 6: Statistical evaluation of the performance of Decision Tree

In the Decision Tree model, excellent training performance is evident with an R^2 of 0.99, suggesting nearly complete variance explanation, accompanied by a low RMSE of 1.35, indicating precise predictions. However, on the testing dataset, it exhibits a lower R^2 of 0.77 and a considerably higher RMSE of 7.37, implying reduced generalization capability and larger prediction errors.



Graph 9: Decission Tree

Descriptive statistics of the data set.

Model	Data	Statistical Parameter			
		R^2	Mean Square Error	Mean Absolute Error	Median Absolute Error
Linear Regression	CBR	0.07	300.25	15.32	12.36
Random Forest	CBR	0.88	26.29	2.21	0.65
Extreme Gradient Boost	CBR	0.91	23.07	2.45	0.85
Gradient Boosting	CBR	0.91	29.9	3.19	1.26
Decision Tree	CBR	0.89	54.35	2.84	0.53

Table. 7: Statistical Information of Dependent and Independent Variables

Conclusion

- Machine Learning models offer superior accuracy in predicting California Bearing Ratio values, increasing the reliability of pavement designs.
- By optimizing resource use, Machine Learning contributes to reduced environmental impact during road construction and maintenance.
- It's crucial to recognize the importance of data quality and model interpretability as challenges to address when adopting Machine Learning for California Bearing Ratio prediction.
- Incorporating uncertainty estimation into CBR predictions can provide engineers with confidence intervals or probabilistic predictions, which are essential for risk assessment in geotechnical engineering projects.
- Various machine learning algorithms, including regression, decision trees, support vector machines, and neural networks, can be used for CBR prediction. Model selection should be based on the specific characteristics of the dataset and the problem at hand.
- The quality and quantity of the training data are crucial for the performance of machine learning models. High-quality, well-labeled datasets with diverse soil samples can significantly enhance prediction accuracy.
- Rigorous model evaluation techniques, such as cross-validation and different performance metrics (e.g., Mean Absolute Error, Root Mean Square Error), should be employed to assess model performance and avoid overfitting.
- Collecting more diverse and comprehensive geotechnical data, including data from different geographic regions and soil types, can improve model generalization and robustness.

References

[1] Janjua, Zohib Shahzad, and Jagdish Chand. "Correlation of CBR with index properties of soil." *International Journal of Civil Engineering and Technology* 7.5 (2016): 57-62.

[2] Patel, Rashmi S., and M. D. Desai. "CBR predicted by index properties for alluvial soils of South Gujarat." *Proceedings of the Indian geotechnical conference, Mumbai*. 2010.

[3] Ahmad, A. T., and A. J. Dhurgham. "Predicting CBR Value from Index Properties of Soils using Expert System." *Global J. of Rese. in Eng.: Civil and Struc. Eng* 11 (2017): 22-28.

[4] Aljanabi, Khalid R. Mahmood, and Nihad Bahaaldeen Salih. "Using Artificial Neural Networks to predict the Unconfined Compressive Strength of Clayey Soils Stabilized by Various Stabilization Agents." *KSCE Journal of Civil Engineering* 27.9 (2023): 3720-3728.

[5] Bharath, A., et al. "Influence and correlation of maximum dry density on soaked & unsoaked CBR of soil." *Materials Today: Proceedings* 47 (2021): 3998-4002.

[6] Ravichandra, A. H., et al. "Prediction of CBR value by using index properties of soil." *Int Res J Eng Technol* 6.7 (2019): 3740-3747.

[7] Krishna, SH Vamsi, B. Sai Santosh, and BHS Sai Prasanth. "Prediction of UCS and CBR of a stabilized Black-cotton soil using artificial intelligence approach: ANN." *Materials Today: Proceedings* (2023).

[8] Lakshmi, S. Muthu, S. Geetha, and M. Selvakumar. "Predicting soaked CBR of SC subgrade from dry density for light and heavy compaction." *Materials Today: Proceedings* 45 (2021): 1664-1670.

[9] Kassa, Semachew Molla, and Betelhem Zewdu Wubineh. "Use of Machine Learning to Predict California Bearing Ratio of Soils." *Advances in Civil Engineering* 2023 (2023).

[10] Likhith, K. M., Vaishnavi Bherde, and Ramu Baadiga3and Umashankar Balunaini. "Prediction of California Bearing Ratio (CBR) of Soils using AI-based Techniques."

[11] Tunbosun, Akinwamide Joshua, et al. "Application of Machine Learning Techniques in Modelling of Soaked and Unsoaked California Bearing Ratio."

[12] Suthar, Manju, and Praveen Aggarwal. "Modeling CBR value using RF and M5P techniques." *Mendel*. Vol. 25. No. 1. 2019.

[13] Taha, S., et al. "Modeling of California bearing ratio using basic engineering properties." *8th International Engineering Conference, At Sharm Al-Sheikh, Egypt. https://www. researchgate. net/publication/305725330_ Modeling_of_California_Bearing_Ratio_using_Basic_ Engineering_Properties*. 2015.

[14] Bhatt, Sudhir, Pradeep K. Jain, and M. Pradesh. "Prediction of California bearing ratio of soils using artificial neural network." *Am. Int. J. Res. Sci. Technol. Eng. Math* 8.2 (2014): 156-161.

[15] Vu, Dung Quang, et al. "Estimation of California bearing ratio of soils using random forest based machine learning." *Journal of Science and Transport Technology* (2021): 48-61.