

Development of a Python-Based Machine Learning Model for California Bearing Ratio (CBR) Prediction

T. Aleyemi¹, V. M. Ishaku², J.V. Madaki³

¹Department of Civil Engineering, Nigerian Defence Academy Postgraduate School, Kaduna, Nigeria.

²Department of Civil Engineering, Federal University Wukari, Taraba, Nigeria.

³Department of Civil Engineering, Shenyang University of Technology, Shenyang, China.

Abstract:

The California Bearing Ratio (CBR) test is the most widely used method for evaluating subgrade strength and for the design of flexible pavements, but the standard soaked CBR procedure typically requires between four and five days to complete, a duration that can be a challenge when infrastructure must be delivered quickly. This study presents a Python-based machine learning approach for the indirect prediction of soaked CBR from index and compaction properties routinely measured in geotechnical practice. Geotechnical laboratory data comprising existing moisture content, liquid limit, plastic limit, plasticity index, percentage passing standard sieve sizes, optimum moisture content, maximum dry density, and specific gravity were compiled for 1,000 soil samples. The dataset was imported into Python using the pandas library and split into training (80%) and testing (20%) subsets. A random forest regression model was then trained to predict soaked CBR values from the remaining variables. On the test set, the model achieved a coefficient of determination (R^2) of 0.957, a mean absolute error of 2.43, and a root mean square error of 3.255, indicating strong predictive performance. These results demonstrate that random-forest-based machine learning models can reliably estimate soaked CBR values without the time and material costs of conventional laboratory soaking, offering a practical complement to standard testing, particularly for time-constrained construction projects.

Keywords: Python; Machine Learning; Random Forest Regression; California Bearing Ratio; Subgrade Soil.

I. INTRODUCTION

The California Bearing Ratio (CBR) test is the most widely used method to evaluate and classify the potential strength of compacted soil subgrade, subbase, base course, or recycled materials for use in road and airfield pavement design and construction. Given the inherent variability of soils as a natural material, different soils will have different CBR values even when compacted with the same amount of energy (Ambrose & Rimoy, 2021). Typical pavement design practice requires representative CBR values from laboratory tests on soil samples taken and soaked for between 48 and 96 hours and compacted to near maximum dry density (MDD). These samples are taken along the proposed alignment in accordance with AASHTO T193 (Ambrose & Rimoy, 2021). This procedure requires relatively large quantities of soil and is time-consuming and costly. As a result, subgrade CBR testing frequencies are practically limited to two points per kilometre, assuming homogeneity or uniform transitions across intermediate sections (Ambrose & Rimoy, 2021).

Various methods have been explored over the years to shorten the time required to obtain CBR results or to make the testing procedure less tedious. Correctly predicting the CBR, or other soil parameters, has therefore become common practice in engineering, and several methods for predicting soil strength have been developed (Eyalet et al., 2024). For example, in predicting the shear strength of soil, cone penetration test (CPT) data and other soil properties have been combined to estimate shear strength parameters (Janibul et al., 2023). Intelligent systems such as artificial neural networks (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) have also been used to predict soil shear strength (Besalatpour et al., 2012; Ashish et al., 2019; Muhammad et al., 2025). These studies generally found ANN models to be more effective than ANFIS models at predicting soil shear strength, with a root mean square error (RMSE), mean estimation error (MEE), and correlation coefficient (R) between measured and ANN-estimated shear strength of 0.05, 0.01, and 0.86, respectively (Muhammad et al., 2025).

Multiple linear regression (MLR), artificial neural networks (ANN), and fuzzy logic (FL) tools have also been used to predict soil strength (Mojtaba, 2012; Aref et al., 2024). For example, Aref et al. (2024) examined variables including the nano-pozzolan content of natural sources (NNP), nano-lime content (NL), median particle size of NNP, active silica content of NNP (SiO_2), and the initial liquid limit (ILL) and initial plastic limit (IPL) of the soils investigated. NNP was added at five percentages (0%, 0.5%, 1%, 1.5%, and 2%), NL was added at five percentages (0%, 0.3%, 0.6%, 0.9%, and 1.2%), and three median particle sizes (50, 100, and 500 nm) were studied (Mojtaba, 2012; Aref et al., 2024; Radha & Smita, 2025; Arif et al., 2019). Based on the soils and combinations investigated, 120 soil mixtures were prepared and tested, with CBR and plasticity index (PI) examined in particular. CBR tests were conducted under soaked conditions on specimens compacted to MDD at the optimum moisture content (OMC), while PI values were obtained from Atterberg limit tests. The results indicated that the CBR and PI of the expansive clayey soils investigated could be effectively predicted using ANN and FL techniques, and that results from MLR were less accurate than those from either ANN or FL; the ANN model was, in turn, found to be slightly more accurate than the FL model for predicting CBR and PI (Aref et al., 2024).

This study makes use of a Python-based machine learning model to predict the CBR value of lateritic soil samples.



II. MATERIALS AND METHODS

2.1 Materials

Geotechnical test results for soil samples, collected over several years, were obtained from a civil engineering laboratory. The dataset comprised existing moisture content (EMC), liquid limit (LL), plastic limit (PL), plasticity index (PI), percentage passing BS sieve Nos. 7, 10, 14, 25, 36, 52, 72, 100, and 200, maximum dry density (MDD), optimum moisture content (OMC), specific gravity (Gs), and unsoaked and soaked CBR values for 1,000 soil samples.

Table 1. Index Properties, Compaction Characteristics, and CBR Values for the Soil Samples

S/no	Depth (m)	EMC (%)	LL (%)	PL (%)	PI (%)	Fines - #200 (%)	MDD (g/cm ³)	OMC (%)	Specific Gravity	CBR Unsoaked (%)	CBR Soaked (%)
1	1	15.8	46	27.7	18	54.4	1.48	18	2.6	28.3	8.7
2	3	16.6	43	32.1	11	86.9	1.6	19.2	2.72	30.3	4.3
3	2.5	22.3	43	27.9	16	64.7	1.76	23.7	2.72	36	9.8
4	1	10.1	0	0	0	8.3	2.06	12	2.59	70.9	46.4
5	2	10.4	0	0	0	16.7	2.01	11.9	2.65	68.2	44.3
6	2	12.4	53	42.4	11	65.9	1.49	15.6	2.75	35.2	7.2
7	2	5.8	0	0	0	21.9	1.83	8.2	2.61	59.6	30.8
8	2.5	10	43	31.2	11	65.4	1.68	13.1	2.68	38.3	10.4
9	2	8	37	24.7	12	80	1.47	12.5	2.7	25.8	2.8
10	1.5	10.4	45	35.1	10	64.7	1.6	15.5	2.69	38.9	9.1
11	1	10.4	33	17.8	15	71.1	1.94	16.5	2.7	47.7	10.4
12	1.5	6.2	0	0	0	26.7	1.9	8.6	2.57	63.9	33.2
13	2	12.8	0	0	0	17.5	2.22	16	2.65	64.4	35.2
14	2	6.5	0	0	0	13.9	2	10.2	2.66	64.3	41
15	2.5	20.1	33	23.7	9.3	84.7	1.65	22.8	2.65	31.2	5.4
16	2	7.3	0	0	0	9.8	1.93	8	2.64	70.7	42.5
17	1	14.8	40	27.8	12	89.6	1.79	21.5	2.76	33.7	9.6
18	3	8.2	0	0	0	11.9	1.85	8.7	2.68	60.8	37.6
19	1	8.4	0	0	0	15.7	1.73	10.4	2.65	51.8	32.7
20	2.5	8.1	0	0	0	11.1	2.1	11.6	2.61	67.4	41.9
21	2.5	12.9	39	24	15	53.7	1.98	14.7	2.8	45.4	13.4
22	1	12.5	20	10	10	60.2	1.67	13.9	2.65	35.1	4.5
23	2	14	67	57.7	9.3	97.6	1.81	21.6	2.68	43.1	5.9
24	1	7.9	0	0	0	13.1	2.01	9.4	2.61	67.5	39.3
25	2	11.2	41	27.6	14	59.2	1.95	18.2	2.8	44.5	8.9
26	3	7.9	0	0	0	22.4	1.85	9.6	2.64	58.1	32.5
27	2.5	11.5	20	10	10	92.1	1.79	15.4	2.71	48.5	12.9
28	1.5	8.3	0	0	0	6.6	2.14	11.7	2.61	73.9	40.3
29	2.5	17.1	39	27.5	11	65.6	1.56	19	2.75	34	7.7
30	2.5	8.3	0	0	0	2.4	1.84	9.5	2.62	58.6	40.6
31	1	16.2	41	22.5	19	59.9	1.45	18.7	2.73	27.6	2
32	3	10.2	0	0	0	2.3	2.02	12.2	2.68	72.3	45.8
33	3	8.5	0	0	0	8.1	1.96	11.3	2.61	64.6	34
34	3	17.6	38	18.8	19	82.7	1.78	19.6	2.74	35.1	9
35	1	12.7	51	43.5	7.6	71.6	1.53	13.8	2.64	36.6	9.1
36	3	7.3	0	0	0	18.7	2	12	2.65	62.2	36.6
37	1.5	8.3	0	0	0	11.9	1.95	10.3	2.64	59.2	36.6
38	1	17.9	40	21.6	18	65.7	1.66	21.8	2.77	29.4	3.6
39	2	6.3	0	0	0	12.1	2.19	7.7	2.65	76.6	43.8
40	1	9.1	0	0	0	10.7	1.9	11.4	2.58	63.8	34
41	1	14.4	35	16.2	19	86	1.65	18.8	2.72	39	6.1
42	1.5	7.1	0	0	0	17.7	2.02	8.8	2.62	62.5	37.2
43	1	8	0	0	0	23.8	2.02	9.7	2.58	68.6	40.8
44	2	13.9	40	20.5	20	98	1.74	17.8	2.7	29.5	3.2
45	3	8.1	0	0	0	19.9	2.11	11.2	2.65	68.4	33.1
46	1	8.4	52	34.1	18	47.6	1.45	9	2.66	40.6	8.4
47	3	12	35	20.6	14	46.6	1.82	18.2	2.7	42.3	12.8
48	2	13.8	36	16.5	20	90	1.87	18.3	2.68	45	5.3
49	2	6.1	22	10	12	76.8	1.9	9.7	2.67	41.3	10.1
50	2.5	15.3	40	30.1	10	67.6	1.85	17.6	2.74	54.2	13.9
51	2.5	12	26	15.1	10	54.5	1.8	16.7	2.7	48.5	14.3



S/no	Depth (m)	EMC (%)	LL (%)	PL (%)	PI (%)	Fines - #200 (%)	MDD (g/cm3)	OMC (%)	Specific Gravity	CBR Unsoaked (%)	CBR Soaked (%)
52	2	8.2	0	0	0	3.4	2.01	8.7	2.62	73.8	42
53	1	17.3	26	13.6	13	35	1.83	22.9	2.65	42.7	17.1
54	3	9.1	0	0	0	29.6	1.83	9.7	2.55	58.6	37.1
55	2.5	13.2	32	20.3	12	89.1	1.85	18.1	2.73	44.7	9.6
56	2.5	19.5	25	10	15	57.9	1.61	21.6	2.71	33.6	6.6
57	2	6.1	0	0	0	10.7	2.14	8.4	2.6	83.1	49.2
58	1.5	10.2	0	0	0	13.5	2.05	11.5	2.58	76.1	44.2
59	2	6.3	0	0	0	19.9	2	9.7	2.55	72.4	46.1
60	1	5.6	0	0	0	16.4	1.94	7.1	2.67	63.2	34.6
61	1.5	8.5	0	0	0	18.9	1.67	12.1	2.6	46.7	25.4
62	3	11.5	0	0	0	18.4	1.96	12.1	2.64	61.5	36.1
63	2	18.4	33	21.1	12	92.5	1.69	19.7	2.72	31.5	2
64	1	5.2	0	0	0	13.4	1.95	8.4	2.66	61	40
65	2	6.6	0	0	0	20.2	2.06	8.4	2.62	63.8	42.3
66	2.5	14.4	49	44.4	4.9	68.6	1.66	20.1	2.63	43.6	14.7
67	1	11.8	44	30.2	13	70.9	1.74	19.3	2.71	32.6	11.9
68	2	14.9	39	21.3	18	42.1	1.75	16.5	2.79	41.3	10.2
69	2.5	13.3	49	39	10	96.8	1.84	17.9	2.74	45	7.3
70	1.5	12.7	0	0	0	10.6	2.18	14	2.57	78.2	47.8
71	1	14	39	29.8	8.8	76	1.75	21.7	2.7	41.5	9.6
72	2.5	6.8	0	0	0	9	2.04	11.3	2.57	72.2	42
73	3	9	0	0	0	8.2	2.02	11.3	2.58	67	45
74	2	6.8	0	0	0	10.4	1.92	9.2	2.63	64.7	36.8
75	2	5	0	0	0	26.5	1.73	6.8	2.57	54.4	22.3
76	2.5	12.7	34	23.2	11	53.1	1.69	20.9	2.66	47.6	17.1
77	2.5	16.7	42	14.9	27	72	1.48	21.9	2.74	23.8	4
78	1	7.4	0	0	0	18.4	2.06	8.5	2.54	69	43.4
79	1.5	10.3	0	0	0	11.7	1.97	11.1	2.6	65.5	42.8
80	1.5	11.6	0	0	0	14.2	2.01	12.4	2.62	61.5	37.5
81	3	15	21	10.3	11	35	1.68	16.7	2.67	40.5	17.9
82	1.5	8.5	0	0	0	15.8	2.17	9.3	2.61	68	38.1
83	1.5	11.3	0	0	0	18.2	1.91	12.5	2.64	57.4	33.7
84	1.5	8	0	0	0	8.1	1.91	10.2	2.62	65.1	41.1
85	1	7.5	0	0	0	18.6	2.18	11.1	2.67	72.2	39.5
500	1.5	10.5	36	28	8.4	82.7	1.8	15.6	2.72	46.2	10.5
501	1.5	14.8	29	23.9	5.2	52.7	1.62	16.2	2.73	44.1	19.2
502	1.5	15.8	41	31.4	10	84.8	1.73	18.2	2.74	39.6	7.8
503	1	10.3	54	30.3	24	49.9	1.45	16.3	2.72	20.1	3.7
504	2	13.7	44	25.7	19	46.5	1.73	16.4	2.7	43.5	10.4
505	2	14.6	33	17.8	15	54.2	1.82	16.6	2.67	44.6	12.5
506	3	10.3	0	0	0	17.6	1.89	11.1	2.61	66.7	37.1
507	1	6.8	0	0	0	4.9	1.99	11	2.61	64	44.7
508	3	7.4	0	0	0	14.1	1.98	11.4	2.62	67.9	45.3
509	3	10.4	0	0	0	18.2	2.06	14.3	2.5	55.9	29.6
510	1.5	18.5	42	25.3	16	87.3	1.45	19.5	2.68	28.2	3.5
511	3	6.8	0	0	0	2.3	2.03	7.6	2.62	67.3	40.8
512	2.5	17.5	42	26.8	15	72.5	1.74	20.3	2.69	39	6.4
513	2	13.3	42	27.7	14	61.7	1.66	15.6	2.68	39.1	10
514	1.5	5.1	0	0	0	12	1.86	7	2.58	64.3	44.3
515	1	10.3	52	35.7	17	66	1.72	16.3	2.8	30.5	7.8
516	1	10.6	0	0	0	7	2.1	11.4	2.63	74.3	58.4
517	2	14.4	36	21.6	14	47.4	1.87	18.5	2.68	49.9	19.3
518	1	12	33	14	19	44.3	1.45	17.9	2.63	28.3	4.7
519	2.5	8.4	0	0	0	20.3	2.12	12	2.55	72.2	40.5
520	3	10.5	0	0	0	9.1	2.03	11.1	2.61	64.1	37.3
521	3	10.4	0	0	0	17.6	2.25	11.6	2.5	74	41.7
522	2	19.7	50	33.7	16	72.2	1.85	25.1	2.7	40.4	8.3
523	2.5	17.3	23	13.4	9.3	65.8	1.75	21.7	2.67	38.5	9.7
524	3	10.3	0	0	0	24	2.03	15	2.58	66.3	36.2



S/no	Depth (m)	EMC (%)	LL (%)	PL (%)	PI (%)	Fines - #200 (%)	MDD (g/cm3)	OMC (%)	Specific Gravity	CBR Unsoaked (%)	CBR Soaked (%)
525	1.5	12.7	35	22.4	13	86.8	1.7	15.2	2.65	44.8	7.6
526	2	9.5	0	0	0	11.7	1.95	11.4	2.63	58.8	34.7
527	2.5	7.9	0	0	0	19.1	2.25	9.7	2.66	76.6	42
528	2	12.6	30	16.4	14	83.9	1.62	16.2	2.63	35.3	6.4
529	1	16	44	29.8	14	91.7	1.87	19	2.75	38.2	6.9
530	1.5	12.4	50	31.2	19	89.7	1.73	16	2.65	33.3	5.7
531	2	13.4	63	47.8	15	72.7	1.51	20.5	2.7	25.9	5.6
532	3	21.7	37	20.3	17	69.2	1.73	23.5	2.78	34.6	7.8
533	2.5	6.8	0	0	0	13	1.98	10.1	2.67	67.5	38.9
534	2.5	13	36	21.8	15	61.7	1.67	17.3	2.79	40	8.9
535	3	9.2	0	0	0	19.6	2.04	11.1	2.61	68.9	40
536	1.5	12.2	25	15.2	10	51.9	1.88	18.3	2.69	46.2	15.3
537	1	6.9	0	0	0	5.7	2.07	11	2.58	64	33.8
538	1.5	15	0	0	0	21.7	2.11	16	2.58	60.6	32.5
539	3	12.2	0	0	0	7.7	1.81	13	2.65	64.8	43.2
540	2.5	8.4	0	0	0	14.3	2.16	11.4	2.62	72.2	38.1
541	1	6.2	0	0	0	10.5	1.81	9	2.6	59.2	30.2
542	1.5	11.6	40	27.8	12	62.8	1.87	17.5	2.6	42.5	14.3
965	3	5.8	0	0	0	4.8	1.95	6.6	2.65	70.3	44.1
966	1	10	38	21.2	17	98	1.78	13.2	2.67	32.6	6.8
967	1	16.1	53	39.3	14	80.8	1.79	17.4	2.75	47.4	7.7
968	2	12.2	33	27.2	5.5	86.9	1.88	15.6	2.67	45.9	9.1
969	3	13.4	43	27.4	16	72.4	1.75	19.7	2.65	39.2	7.4
970	1.5	4.4	0	0	0	14	1.84	6.3	2.58	60.7	31.7
971	2.5	13.9	23	14.6	8.8	72.2	1.73	15.8	2.64	40.6	8.7
972	2.5	9.3	0	0	0	20.2	1.76	10.4	2.61	53.4	30.2
973	2	17.7	44	29.1	15	71.1	1.51	18.8	2.72	30.5	5.6
974	1	14.1	53	39.1	14	80.3	1.85	15.6	2.71	45.1	6
975	2.5	19.6	22	10	12	91.5	1.67	22.3	2.6	34.2	2.6
976	1.5	5.8	0	0	0	29.9	1.93	8.6	2.57	62.3	34
977	2.5	8.9	0	0	0	11.2	1.95	9.8	2.57	65.6	41.3
978	3	9.9	0	0	0	6.6	1.85	12.6	2.58	68.5	43.8
979	3	12.2	49	38	11	61.2	1.66	15.2	2.77	37.8	14.8
980	1.5	8.6	0	0	0	21.3	1.9	9.6	2.64	58.1	33.2
981	1	13.8	40	20.9	19	74.6	2.02	17.6	2.75	41.8	6.4
982	2	9	0	0	0	5.9	2	10.8	2.63	75.4	52.5
983	2.5	5.2	0	0	0	20.3	1.92	7.9	2.55	55.4	31.8
984	3	14.5	33	13.1	20	52.3	1.51	17.2	2.74	22	7.2
985	1	15.1	39	16.5	23	68.5	1.79	16.6	2.7	33.5	6.4
986	2.5	9.1	0	0	0	20.7	2.11	10.6	2.59	68.3	40.1
987	1.5	16.1	26	12.4	14	86.9	1.58	19.9	2.76	31.6	3.5
988	3	6.9	0	0	0	17.7	1.98	10.8	2.6	61.9	40.5
989	1	20.2	33	16.9	16	65.1	1.45	24.2	2.8	26.5	8.6
990	2	13.8	39	21.4	17	40.9	1.85	21	2.69	50.7	14
991	1.5	14.9	20	10	10	71	1.99	21.3	2.63	46.3	15.8
992	1	11.8	53	32.4	21	85.5	1.82	18.3	2.75	39.4	5.6
993	2.5	10.5	56	34.9	22	55	1.68	15.3	2.7	29	5.2
994	1.5	7.5	0	0	0	31.3	2.06	10.7	2.61	69.2	34.2
995	1.5	9.2	30	14.3	15	63.2	1.84	14.7	2.72	50.6	11.3
996	2	6.1	0	0	0	24.4	1.98	8.8	2.66	62.9	29.5
997	1.5	10.8	0	0	0	26.8	2.05	12.1	2.58	71.9	37.1
998	2.5	15.6	41	23.3	18	54.1	1.59	17.2	2.71	34	6
999	3	15.3	40	21.2	19	61.6	1.93	18.2	2.76	39.4	4.7
1000	2	19.2	29	13.7	16	74.6	1.55	23.7	2.67	29.7	5

Note: A representative subset of the dataset is shown above; the complete dataset ($N = 1,000$) was used in the analysis.



2.2 Methods

2.2.1 Python-Based Machine Learning Modelling

The Excel file containing the soil sample data was imported into the Python environment using the pandas library. The dataset was displayed using the print command, as shown in Figure 1. By default, Python displays only the first five rows of a data frame when the print command is issued, even though the remaining rows are still held in memory.

```
##LOAD SOIL TEST RESULT
import pandas as pd
df = pd.read_excel(r"c:\Users\HP USER\Desktop\soil samples results from lab.xlsx")
print(df.head())
```

[1] ✓ 7.1s

S/no	Depth (m)	EMC (%)	LL (%)	PL (%)	PI (%)	Fines - #200 (%)	MDD (g/cm3)	OMC (%)	Specific Gravity	CBR Unsoaked (%)	CBR Soaked (%)
0	1	15.8	45.5	27.7	17.8	54.4	1.48	18.0	2.60	28.3	8.7
1	2	16.6	42.7	32.1	10.6	86.9	1.60	19.2	2.72	30.3	4.3
2	3	22.3	43.3	27.9	15.5	64.7	1.76	23.7	2.72	36.0	9.8
3	4	10.1	0.0	0.0	0.0	8.3	2.06	12.0	2.59	70.9	46.4
4	5	10.4	0.0	0.0	0.0	16.7	2.01	11.9	2.65	68.2	44.3

Figure 1. Loading of soil test result.

To view the complete dataset, a command was issued to print a data frame of 1,010 rows. Because the dataset contained 1,000 soil samples, with each row representing one sample, printing a figure higher than 1,000 rows allowed both the first and last rows to be checked in the output, confirming that the full dataset had been captured, as shown in Figure 2.

```
## PREVIEW FULL DATASET
print(df.head(1010))
```

[3] ✓ 0.0s

S/no	Depth (m)	EMC (%)	LL (%)	PL (%)	PI (%)	Fines - #200 (%)	MDD (g/cm3)	OMC (%)	Specific Gravity	CBR Unsoaked (%)	CBR Soaked (%)
0	1	15.8	45.5	27.7	17.8	54.4	1.48	18.0	2.60	28.3	8.7
1	2	16.6	42.7	32.1	10.6	86.9	1.60	19.2	2.72	30.3	4.3
2	3	22.3	43.3	27.9	15.5	64.7	1.76	23.7	2.72	36.0	9.8
3	4	10.1	0.0	0.0	0.0	8.3	2.06	12.0	2.59	70.9	46.4
4	5	10.4	0.0	0.0	0.0	16.7	2.01	11.9	2.65	68.2	44.3
...	...	...	...	...	...	...	...	...	...	...	...
995	996	2.0	6.1	0.0	0.0	24.4	1.98	8.8	2.66	62.9	29.5
996	997	1.5	10.8	0.0	0.0	26.8	2.05	12.1	2.58	71.9	37.1
997	998	2.5	15.6	41.1	23.3	54.1	1.59	17.2	2.71	34.0	6.0
998	999	3.0	15.3	39.7	21.2	61.6	1.93	18.2	2.76	39.4	4.7
999	1000	2.0	19.2	29.4	13.7	74.6	1.55	23.7	2.67	29.7	5.0

[1000 rows x 12 columns]

Figure 2. Output confirming that all 1,000 soil samples were captured in the data frame.

The data was then split into two sets: a training set, accounting for 80% of the data (800 samples), and a test set, accounting for the remaining 20% (200 samples), as shown in Figure 3.

```

▶ ✓ ##SPLIT DATA INTO TRAINING AND TEST SETS
    from sklearn.model_selection import train_test_split
    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
    print(x_train.shape, x_test.shape)

[6] ✓ 4.1s

... (800, 11) (200, 11)

```

Figure 3. Splitting of the dataset into training and test sets.

The random forest regressor, which is the algorithm used to generate the prediction, was then initialised and fitted to the training data, as shown in Figure 4.

```

▶ ✓ ## INITIALIZE THE RANDOM FOREST MODEL
    from sklearn.ensemble import RandomForestRegressor
    model = RandomForestRegressor(random_state = 42)

[8] ✓ 1.2s

▶ ✓ ## TRAIN THE RANDOM FOREST MODEL
    model.fit(x_train, y_train)

[10] ✓ 0.7s

...
    RandomForestRegressor ⓘ ?
    ▶ Parameters
    ▶ Fitted attributes

```

Figure 4. Initializing the random forest model.

After the model was fitted and the print command issued, the predicted values for the 20% of the data reserved for testing were displayed, as shown in Figure 5.

```

D v
## GENERATE PREDICTIONS ON TEST DATA
y_pred = model.predict(x_test)
print (y_pred)
[11] ✓ 0.0s

... [ 8.376 47.215  6.231 35.016 40.58   7.643 42.742 39.533 40.464 40.565
12.486  4.164  8.214  7.958  9.4    7.839  3.249  6.409  8.784 48.55
43.799 41.226  7.76   6.356  4.504 33.652 41.032  5.639 48.178 11.475
 6.56  28.944 38.936  9.266  5.817 40.705  4.405  6.319 36.946 39.455
 7.051 37.488 38.133  7.059  9.404  5.929 36.127 31.25 38.373  3.274
34.803  9.604  6.952 43.681 42.211  6.047  6.061  9.68 44.967  8.334
45.624  7.782 30.622 40.709 10.07  43.55 41.031 37.913  9.4  37.964
 8.884  5.12  33.586  4.812 16.394 11.175  9.642 37.649 44.693 31.921
 8.463  7.584 39.074 11.488 36.03  36.01  7.946 38.441 29.807  9.189
34.92   7.47   9.213 16.76   6.578 11.269 10.668 33.172 35.39 10.845
41.264  5.464  9.842  8.769 11.307 43.767  6.192  7.79  36.61  6.088
 6.378 38.247 38.383 33.429 34.607 30.752  6.442 10.727 40.373  8.582
10.379 42.751 34.914 34.146 44.96   9.658 33.186 34.202  6.582  7.779
 9.162 37.399 34.471 35.727  6.485  8.038 43.54  40.207 32.541 39.681
40.933  8.334 47.38  44.366  4.867  3.448 42.563 36.195  6.198  6.887
 7.418 10.138 44.531  7.903 42.116 32.375  9.249 37.705 27.639  9.95
 8.928 37.661 16.257 34.692 38.722  6.059 42.924 10.028  3.835  7.839
44.657 13.48   8.772 15.401 37.156 11.405 33.886  3.213 36.144 17.624
 6.515  5.054 31.174 45.864 13.711 12.833  8.177 40.387 41.192  5.516
 9.312  8.022 39.298  4.825 13.583 37.369  7.507  5.707 32.996 39.13 ]

```

Figure 5. Generating the test-set predictions.

III. RESULTS AND DISCUSSION

The coefficient of determination (R^2) of the prediction model was found to be 0.957, which is close to 1, the value for a perfect prediction. The mean absolute error (MAE) was found to be 2.43, meaning that the predicted CBR values deviated from the true values by an average of only 2.43. This is a strong result given that BS 1377 stipulates that the soaked CBR value for sub-base course should not be less than 30, and the unsoaked CBR value for base course should not be less than 80; an MAE of this size is therefore low relative to the required values. The root mean square error (RMSE) of 3.255 was also low, further indicating a strong prediction (Figure 6).

```

D v
## EVALUATE MODEL PERFORMANCE
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
import numpy as np

r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))

print("R² score:", r2)
print("MAE:", mae)
print("RMSE:", rmse)
[12] ✓ 0.0s

... R² score: 0.9574735902806921
MAE: 2.431415
RMSE: 3.2557074615204606

```

Figure 6. Evaluating model performance.

A comparison was then carried out between the actual values obtained from the laboratory and the values predicted by the trained Python model, with the resulting error clearly shown below (Figures 7a and 7b).

```

##COMPARISON TABLE WITH DIFFERENCE / Abs DIFFERENCE / % ERROR
import pandas as pd

# Build the comparison table
comparison = pd.DataFrame({
    "Actual": y_test,
    "Predicted": y_pred
})

comparison["Difference"] = comparison["Actual"] - comparison["Predicted"]
comparison["Abs Difference"] = comparison["Difference"].abs()
comparison["% Error"] = (comparison["Difference"] / comparison["Actual"]) * 100

print(comparison)

```

[13] ✓ 0.0s

Figure 7a. Python command used to compare the actual and predicted CBR values.

...	Actual	Predicted	Difference	Abs Difference	% Error
521	8.3	8.376	-0.076	0.076	-0.915663
737	45.0	47.215	-2.215	2.215	-4.922222
740	7.1	6.231	0.869	0.869	12.239437
660	31.9	35.016	-3.116	3.116	-9.768025
411	34.5	40.580	-6.080	6.080	-17.623188
..	...	...	...	...	...
408	38.8	37.369	1.431	1.431	3.688144
332	7.0	7.507	-0.507	0.507	-7.242857
208	4.5	5.707	-1.207	1.207	-26.822222
613	39.0	32.996	6.004	6.004	15.394872
78	42.8	39.130	3.670	3.670	8.574766

[200 rows x 5 columns]

Figure 7b. Output comparing the actual and predicted CBR values.

The predicted soaked CBR values were then compared with the actual CBR values obtained from the laboratory (Figures 8a and 8b, and Figures 9a and 9b).

```

## BAR CHART OF PREDICTED ERROR
import matplotlib.pyplot as plt
import numpy as np

errors = y_test.values - y_pred

plt.figure(figsize=(10,6))
plt.bar(range(len(errors)), errors, color='tomato')
plt.axhline(0, color='black', linewidth=0.8)
plt.xlabel("Test Sample Index")
plt.ylabel("Error (Actual - Predicted)")
plt.title("Prediction Errors per Sample")
plt.show()

```

[17] ✓ 1.0s

Figure 8a. Code used to print the predicted error.

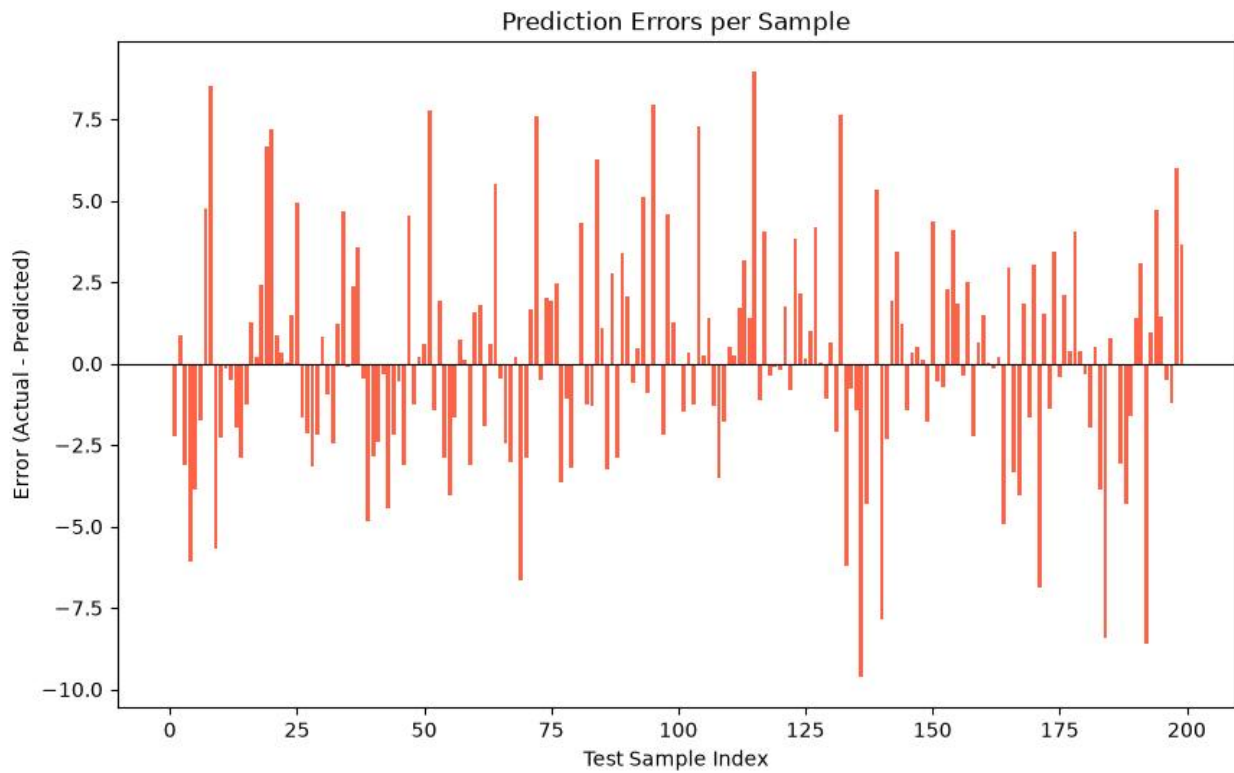


Figure 8b. Predicted error per sample.

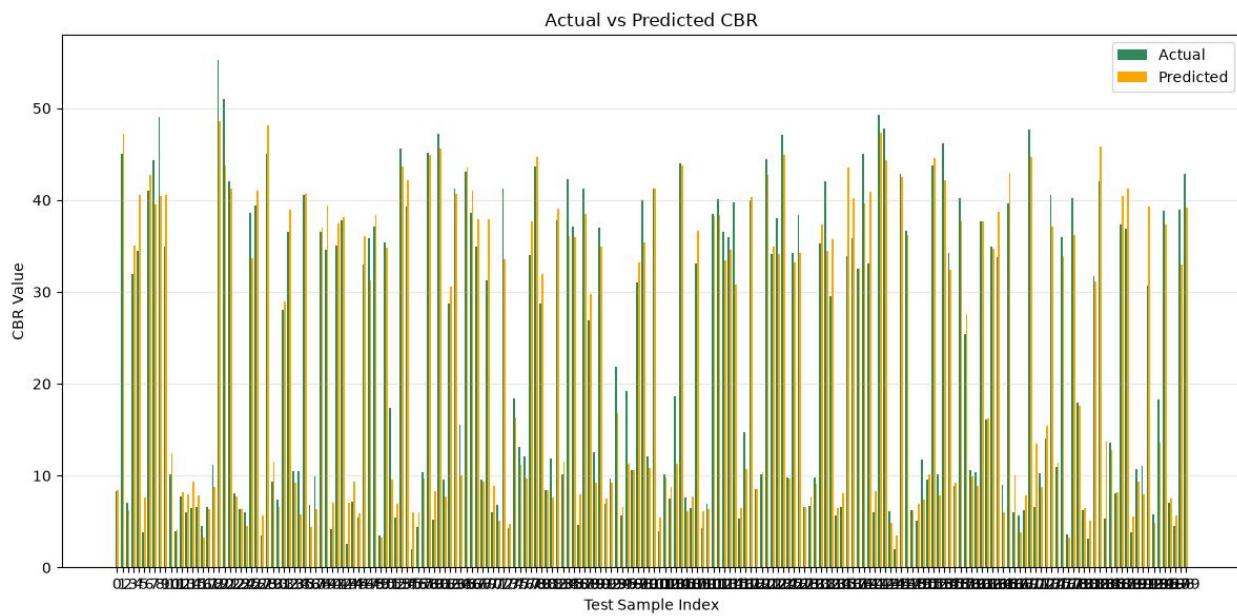


Figure 9a. Bar chart comparing the actual and predicted CBR values.

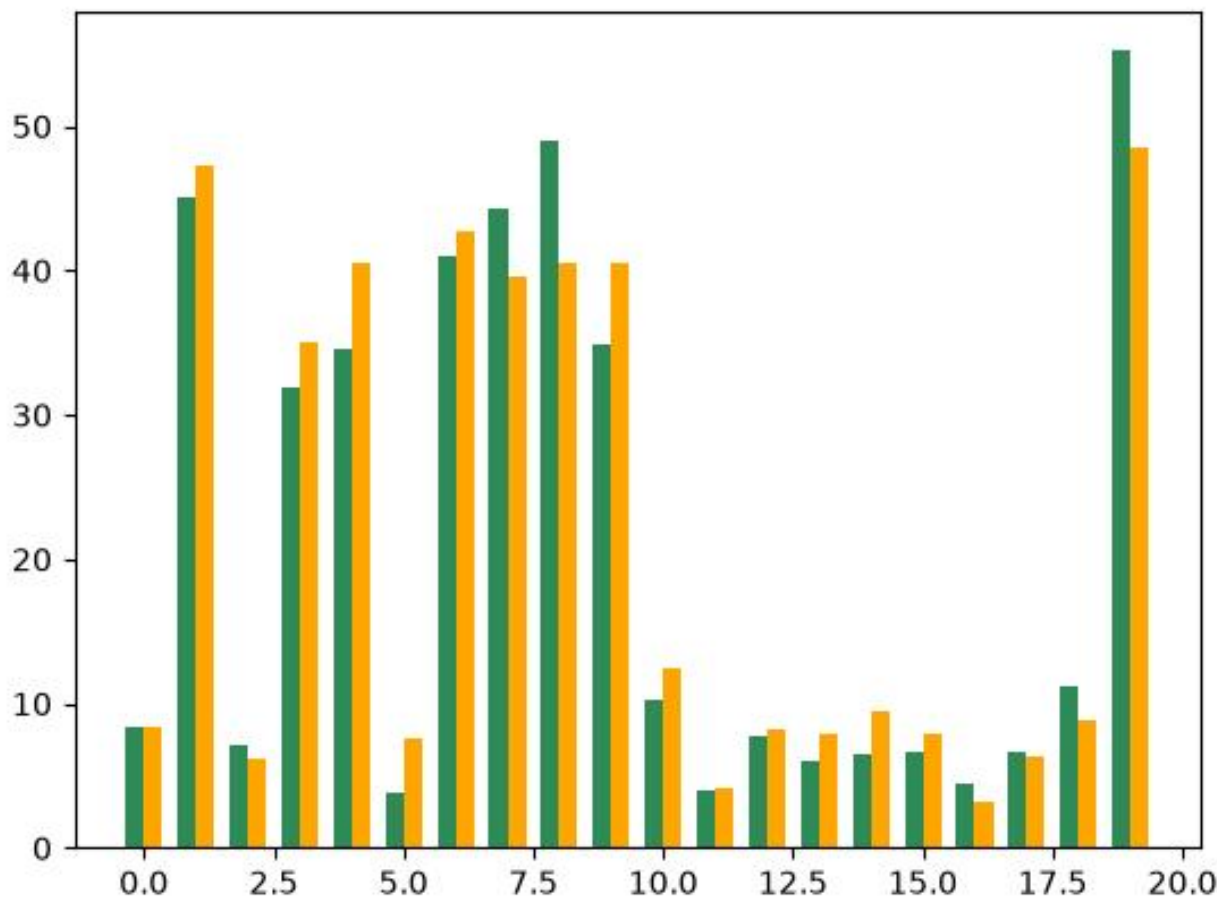


Figure 9b. Bar chart comparing the actual and predicted CBR values for 20 random samples taken from the test set.

IV. CONCLUSION

This study developed and evaluated a Python-based machine learning approach for the indirect prediction of the soaked California Bearing Ratio (CBR) of subgrade soils, using index and compaction properties routinely measured in geotechnical practice. Based on the analysis of 1,000 soil samples, the following conclusions are drawn:

1. It is possible to determine the soaked CBR of a soil by an indirect machine learning method, with a high degree of accuracy.
2. Using 1,000 soil samples, with 80% used for training and 20% for testing, the model was able to predict the CBR value of soil samples with a coefficient of determination (R^2) of 0.957, a mean absolute error (MAE) of 2.43, and a root mean square error (RMSE) of 3.255, all of which are strong indications of good predictive performance.

These findings support the continued development of indirect, machine-learning-based methods for CBR estimation as a practical complement to conventional laboratory testing, particularly where time constraints make the standard four- to five-day soaked CBR procedure impractical. Future work using a larger and more geographically diverse sample set is recommended to further improve and validate the general ability of the predictive model.

REFERENCES

- Aderinola, O. S. (2017). Predicting the California bearing ratio value of low compressible clays with its index and compaction characteristics. *International Journal of Scientific & Engineering Research*, 8(5), 1460–1472.
- Alawi, M. H., & Rajab, M. I. (2013). Prediction of California bearing ratio of subbase layer using multiple linear regression models. *Road Materials and Pavement Design*, 14(1), 211–219.
- Baghbani, A., Nguyen, M. D., Alnedawi, A., Milne, N., Baumgartl, T., & Abuel-Naga, H. (2023). Improving soil stability with alum sludge: An AI-enabled approach for accurate prediction of California bearing ratio. *Applied Sciences*, 13(8), 4934.
- Bardhan, A., Gokceoglu, C., Burman, A., Samui, P., & Asteris, P. G. (2021). Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. *Engineering Geology*, 291, 106239.
- Bayamack, J. F. N., Onana, V. L., Mvindi, A. T. N., Ze, A. N. O., Ohandja, H. N., & Eko, R. M. (2019). Assessment of the determination of Californian Bearing Ratio of laterites with contrasted geotechnical properties from simple physical parameters. *Transportation Geotechnics*, 19, 84–95.
- Black, P. B. (1961). Design of flexible pavements using the California Bearing Ratio method. Highway Research Board.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Chu, X., Dawson, A., & Thom, N. (2021). Prediction of resilient modulus with consistency index for fine-grained soils. *Transportation Geotechnics*, 31, 100650.
- Das, B. M. (2016). *Principles of Geotechnical Engineering* (9th ed.). Cengage Learning.



- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- González Farias, I., Araujo, W., & Ruiz, G. (2018). Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. *Geotechnical and Geological Engineering*, 36(6), 3485–3498.
- Ho, L. S., & Tran, V. Q. (2022). Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste. *Journal of Cleaner Production*, 370, 133587.
- Kassa, S. M., & Wubineh, B. Z. (2023). Use of machine learning to predict California bearing ratio of soils. *Advances in Civil Engineering*, 2023, Article 8198648.
- Kim, M., Ordu, S., Arslan, O., & Ko, J. (2023). Prediction of California bearing ratio (CBR) for coarse- and fine-grained soils using the GMDH-model. *Geomechanics and Engineering*, 33(2), 183–195.
- 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.
- Kurnaz, T. F., & Kaya, Y. (2019). Prediction of the California bearing ratio (CBR) of compacted soils by using a GMDH-type neural network. *European Physical Journal Plus*, 134, 326.
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.
- Othman, K., & Abdelwahab, H. (2023). The application of deep neural networks for the prediction of California Bearing Ratio of road subgrade soil. *Ain Shams Engineering Journal*, 14, 101988.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rakaraddi, P. G., & Gomarsi, V. (2015). Establishing the relationship between CBR with different soil parameters. *International Journal of Research in Engineering and Technology*, 4(2), 182–188.
- Rehman, A., Farooq, K., & Mujtaba, H. (2017). Prediction of California bearing ratio (CBR) and compaction characteristics of granular soils. *Acta Geotechnica Slovenica*, 14(1), 62–72.
- Shirur, N. B., & Hiremath, S. G. (2014). Establishing relationship between CBR value and physical properties of soil. *IOSR Journal of Mechanical and Civil Engineering*, 11(5), 26–30.
- Suthar, M., & Aggarwal, P. (2018). Predicting CBR value of stabilized pond ash with lime and lime sludge using ANN and MR models. *International Journal of Geosynthetics and Ground Engineering*, 4, 1–7.
- Taha, S., Gabr, A., & El-Badawy, S. (2019). Regression and neural network models for California bearing ratio prediction of typical granular materials in Egypt. *Arabian Journal for Science and Engineering*, 44(10), 8691–8705.
- Taskiran, T. (2010). Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Advances in Engineering Software*, 41(6), 886–892.
- Tenpe, A. R., & Patel, A. (2020). Application of genetic expression programming and artificial neural network for prediction of CBR. *Road Materials and Pavement Design*, 21(5), 1183–1200.
- Yildirim, B., & Gunaydin, O. (2011). Estimation of California bearing ratio by using soft computing systems. *Expert Systems with Applications*, 38(5), 6381–6391.

