Graduate Journal of Interdisciplinary Research, Reports & Reviews
(Grad.J.InteR^3)
ark:/85846/JPR3.2024AR.0102054

ISSN(E): 2584-2919
Vol.1, No.2 (2024)
pp.54-62

Research Article

Analysis in Materials Science by Predicting Concrete Compressive Strength Using Machine Learning

T. Hakimi*, M. Bhuyan**

Center for Theoretical and Computational Physics, Department of Physics, Faculty of Science, Universiti Malaya, Kuala Lumpur 50603, Malaysia.

`Email:`taufiq_hakimi@yahoo.com (corresponding author)

**`ORCID:`0000-0002-8677-4220 (M Bhuyan)*

(Received: 26-02-24; Accepted:12-05-24)

ABSTRACT: Future developments in materials science engineering will be greatly influenced by the application of machine learning for determining the properties of concrete, especially its compressive strength. This research predicts the compressive strength of concrete with eight independent variables, including cement, blast furnace slag, fly ash, water, super plasticizers, coarse aggregate, fine aggregate, and age using supervised machine learning (ML) techniques of linear regression (LR) and light gradient boosting machine (LGBM). The ML models are fed a total of 1030 data-sets using a 70:30 split ratio for training and testing. Performance metrics like $R2$, MAE, MSE, and RMSE are used to assess how well the ML models are in making predictions. From the research, the LR model ($R2$ value of 0.607) is less effective than the LGBM model ($R2$ value of 0.920) in predicting compressive strength. Furthermore, feature importance predicted by LGBM shows that the cement content (2331), fine aggregate (2200), and coarse aggregate (2076) all significantly influence the prediction of concrete compressive strength.

Keywords: Material Science, Machine learning, Properties of concrete, Cement, Fly ash, super-plasticizers

1. Introduction

In the field of materials science, machine learning (ML) is widely used. ML is a branch of Artificial Intelligence (AI) that uses algorithms to self-learn and enhance its performance using past data-sets. ML algorithms will automatically learn and get better over time with very little human involvement [1]. AI and ML have already been used in engineering to overcome issues in various structural engineering domains [2]. Further applications of machine learning include the prediction and evaluation of concrete characteristics, the improvement of finite element modelling of buildings, and building structural design and performance assessment [3].

Compressive strength is the most important of the several concrete properties, as it is used to evaluate the performance of structures, from new structural design to old structural assessment [4]. Cement, water, fine aggregate, and coarse aggregate are the four main ingredients of concrete [5] . To improve the quality of concrete, additional materials such as industrial wastes or by-products are occasionally added [6]. Each of the ingredients has its unique properties that contribute to the overall strength of the concrete. Cement has a significant impact on the most critical elements of a concrete mixture, including workability, compressive strength, drying shrinkage, and durability. Water starts the cement's hydration process and gives the mixture workability. The ratio of water to cement is important because too little water can make the concrete difficult to work with and too much water can weaken it [7]. Fine aggregates are typically made up of natural sand or broken stone, with the majority of the particles going through a 3/8-inch screen. Strength

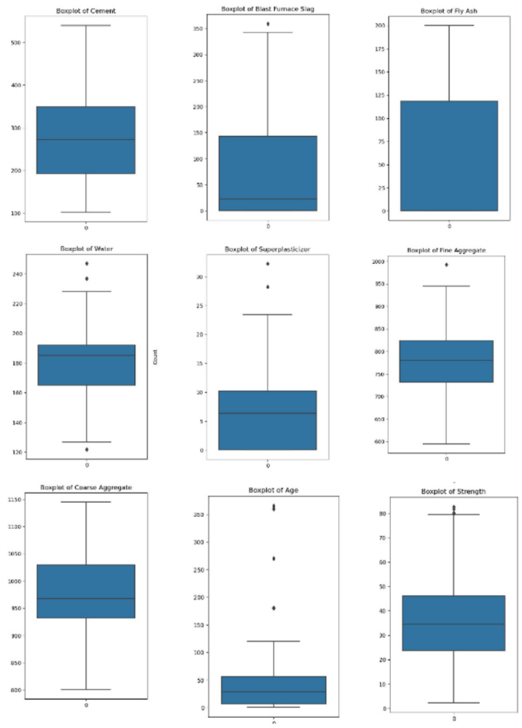

Figure 1: Boxplot of nine variables

is increased because it fills up the voids between cement and coarse particles, providing better particle packing. Coarse aggregates are any particles larger than 0.19 inches, nevertheless, their typical diameter ranges from 3/8 to 1.5 inches. It gives concrete construction strength, thermal and elastic properties, as well as dimension and volume stability. Fly ash is an additional cementitious substance that decreases permeability and increases workability and is produced as a byproduct of burning coal [8]. Blast furnace slag is a byproduct produced during the creation of pig iron, or iron and steel. The benefit of it is being able to replace more cement, allowing for up to 9% in cement cost reductions when the ratio of slag to cement content is 50% [9]. Super plasticizers are polymeric dispersant that are utilized in cementitious materials. It enhances workability without adding more water and lowers the amount of water needed to reach a specific workability level [10]. Age refers to the period that passes after the concrete is placed. Continuous hydration causes concrete to get stronger over time. With machine learning becoming more and more popular, several research was conducted to utilize ML techniques. To predict the concrete compressive strength, various empirical and statistical models like linear and nonlinear regression algorithms were used [11]. Several types of concrete, including conventional [12], high-performance [13], ultra-high performance [3], and green concrete [14] with additional cementitious materials including fly ash, blast furnace slag, and recycled aggregates, were generally done to predict their compressive strength. Other studies have successfully predicted high-performance concrete (HPC) containing silica nanoparticles and copper slag compressive strength using the artificial neural network (ANN) algorithm [15]. Similarly, with the previous, it forecasted the compressive strength of concrete, achieving a value of R2 as high as 0.90 [16]. Nguyen [17] obtained good output accuracy using gradient boosting regressor (GBR) and extreme gradient boosting (XGBoost) to predict the compressive and tensile strength of HPC, although he used four machine-learning algorithms in his research. Kumar [18] predicted the compressive strength of lightweight concrete (LWC) by introducing several machine-learning algorithms, and the best is the support vector machine (SVM) model. Zhang [19] predicted the lightweight self-compacting concrete uni-axial compressive strength while also performing analysis on eight input variables such as characteristic importance, by using random forest (RF).

In this study, machine learning is used for predicting concrete compressive strength. The data set used in this paper was acquired in an experiment conducted at the Chinese University of Taiwan by Professor Yi-Zheng Yeh and his group. The data set was donated to the University of California, Irvine's Machine Learning Laboratory for free. Professor Yi-Zheng Yeh and his colleagues measured the compressive strength of concrete by creating 150 mm-tall cylindrical concrete specimens and testing them using traditional compressive methods following a standard curing period in which the eight variables were gathered [20]. To choose and prepare important characteristics for model input, feature engineering is applied. The job complexity is then taken into consideration using machine learning models and techniques, including linear regression and Light Gradient Boosting Machine (LGBM). To predict concrete strength, the model first learns patterns and correlations from past data during the training phase.

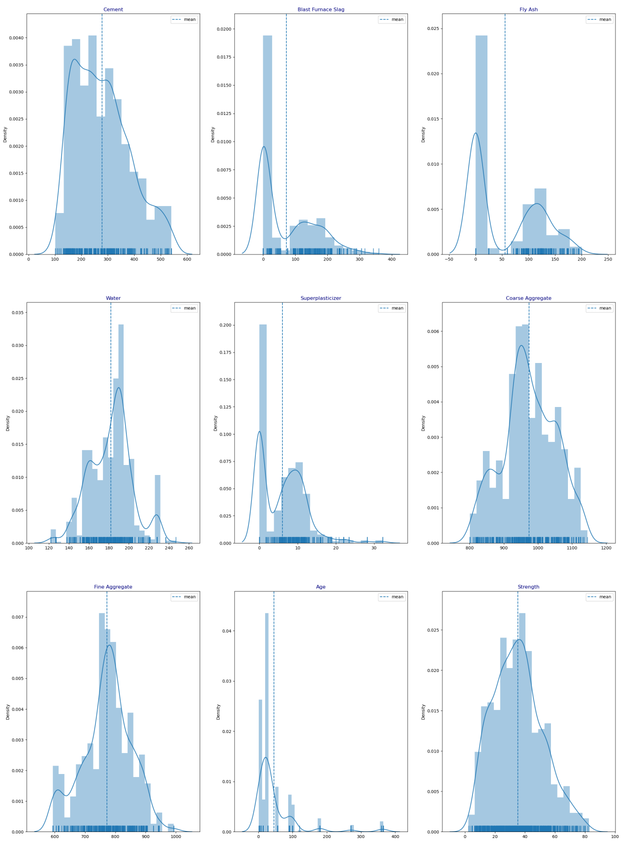

Figure 2: Histogram distribution plot of nine

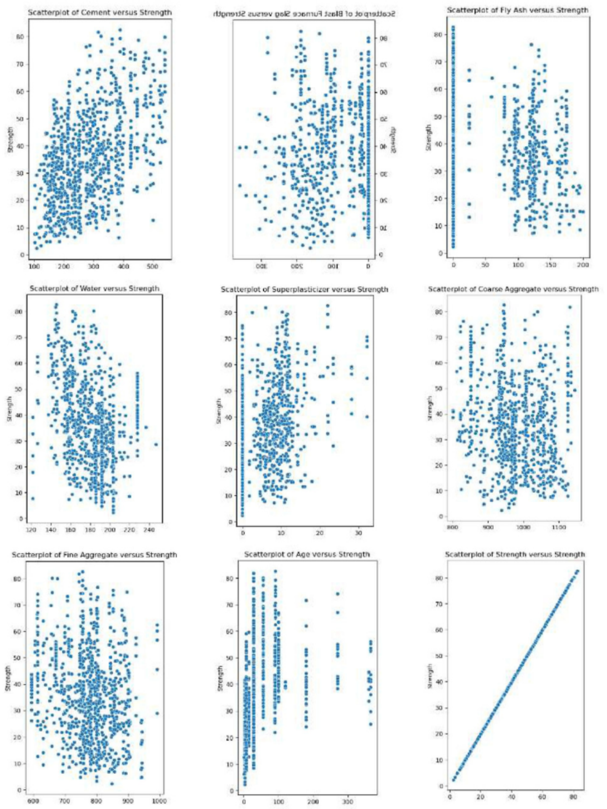

Figure 3: Scatterplot of nine variables.

Metrics like R-Two, Mean Square Error, Root Mean Square Error, and Mean Absolute Error are employed for assessment during training and testing to measure the model performance on untested data. The trained model may be then used to provide future predictions in real time. This helps make better decisions and improve efficiency by incorporating machine learning to predict concrete strength.

2. Method & Materials

2.1. Datasets

The datasets that will be used to conduct this study are shown below in Table 1. Table 1 shows a sample dataset used for this research. The dataset contains 9 variables, where 8 of them (cement, blast furnace slag, fly ash, water, superplasticiser, coarse aggregate, fine aggregate, age) are the quantitative input variable (given unit: kg/m^3) and (given unit: days) used to predict the quantitative output variable, compressive strength (given unit: MPa, megapascals).

2.2. Exploratory Data Analysis (EDA)

To understand this dataset more intuitively, descriptive statistics were performed to gain more insight into the dataset. The following Table 2 shows summary statistics from the datasets. Figure 1 shows the plotting of box plots of the nine variables in the datasets. For each graph, the straight line on top is the value of maximum, the straight line at the bottom is the value of minimum, and the horizontal line in each box is the value of median. Diamond-shaped represents the outliers. Figure 2 shows the plotting of histogram for each variable, while also reflecting the distribution for the attributes and fitting the corresponding normal distribution curves. Figure 3 shows the scatterplot graph for each attribute corresponding to the concrete compressive strength. Figure 4 is the Pearson heatmap coefficient. The correlation coefficient between two variables is represented by each cell in the heatmap, with colour intensity indicating the correlation's strength and direction. Colours that indicate positive correlations are blue, while those that indicate negative correlations are red.

2.3. Machine Learning Approach

2.3.1. Linear Regression (LR)

A more advanced from the simple regression model, linear regression (LR) finds out the correlation between two or more explanatory variables and a

Table 1: A sample dataset (first 10 rows)

Cement	Blast Furnace Slag	Fly Ash Water	Water	Super plasticizers	Coarse Aggregate	Fine Aggregate	Age	Strength
540	0	0	162	2.5	1040	676	28	79.99
540	0	0	162	2.5	1055	676	28	61.89
332.5	142.5	0	228	0	932	594	270	40.27
332.5	142.5	0	228	0	932	594	365	41.05
198.6	132.4	0	192	0	978.4	825.5	360	44.3
266	114	0	228	0	932	670	90	47.03
380	95	0	228	0	932	594	365	43.7
380	95	0	228	0	932	594	28	36.45
266	114	0	228	0	932	670	28	45.85
475	0	0	228	0	932	594	28	39.29

Table 2: Summary statistics

	count	mean	std	min	25%	50%	75%	max
Cement	1030.0	281.167864	104.506364	102.00	192.375	272.900	350.000	540.00
Blast Furnace Slag	1030.0	73.895825	86.279342	0.00	0.000	22.000	142.950	359.4
Fly Ash	1030.0	54.188350	63.997004	0.00	0.000	0.000	118.300	200.1
Water	1030.0	181.567282	21.354219	121.80	164.900	185.000	192.000	247.0
Superplasticizer	1030.0	6.204660	5.973841	0.00	0.000	6.400	10.200	32.2
Coarse Aggregate	1030.0	972.918932	77.753954	801.00	932.000	968.000	1029.400	1145.0
Fine Aggregate	1030.0	773.580485	80.175980	594.00	730.950	779.500	824.000	992.6
Age	1030.0	45.662136	63.169912	1.00	7.000	28.000	56.000	365.0
Strength	1030.0	35.817961	16.705742	2.33	23.710	34.445	46.135	82.6

numerical response variable. This study investigated how applicable LR is. The general least square is given with a problem with n inputs (or independent variables), X's, and one output (or dependent variable), Y, to find out the unknown parameters, as shown in Figure 5.

2.3.2. Light Gradient Boosting Machine (LGBM)

Gradient boosting framework LGBM primarily uses tree-based learning algorithms. Unlike other boosting algorithms that grow the tree level-wise, LGBM splits the tree leaf-wise. Trees grow vertically in the case of LGBM while growing horizontally in other algorithms. To grow, it selects for growth the leaf that has the greatest delta loss. The loss of the leaf-wise algorithm is less than that of the level-wise algorithm because the leaf is fixed. In small data-sets, leaf-wise tree growth may cause overfitting and raise the model's complexity.

2.3.3. Performance measurement

The performance measurement model used to calculate each model's accuracy is displayed in the equations below. The coefficient of determination (R^2), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) are the four performance measurement models. These were introduced to accurately assess how well LR and LGBM are in predicting the compressive strength of concrete. The accuracy of the model is determined by comparing the actual data with the output variable's predicted results.

2.3.4. Coefficient of determination (R^2)

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad (1)$$

where, SS_{res} = sum of squared residuals, SS_{total} = total sum of squares, $\hat{y}_i$ = mean of all the concrete compressive strength.

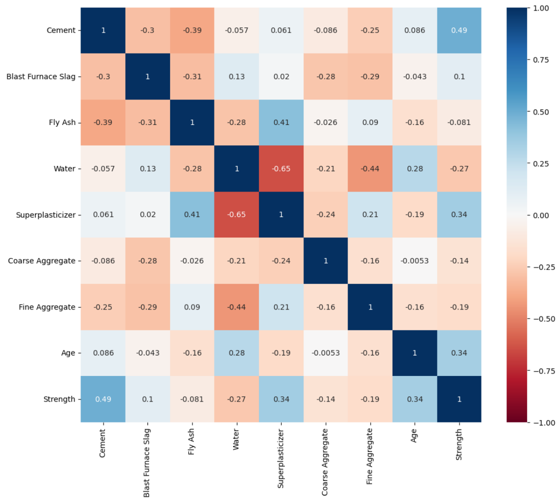

Figure 4: Pearson heatmap correlation.

2.3.5. *Mean Absolute Error (MAE)*

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

n = the total number of samples in the dataset, y_i = actual value of concrete compressive strength, $\hat{y}_i$ = the predicted value of concrete compressive strength.

2.3.6. *Mean Square Error (MSE)*

$$MSE = \frac{1}{n}\sum_{(i=1)}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

2.3.7. *Root Mean Square Error (RMSE)*

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{(i=1)}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

3. Results and Discussions

3.1. *Performance measurement for both model*

Table 3 below displays the testing results for LR and LGBM models. LGBM has been able to outperform LR in each of the four accuracy metrics. With an R^2 value of 0.920, the LGBM model set higher than the LR model and approaches the value of 1. In the meantime, LGBM generates lower errors compared to LR. This overall shows that the LGBM model gives higher prediction accuracy and maintains lower errors than the LR model.

Table 3: Testing results for LR and LGBM

	Testing Results			
Models	**R^2**	**MAE**	**MSE**	**RMSE**
LR	0.607	8.208	106.56	10.32
LGBM	0.920	2.992	21.61	4.649

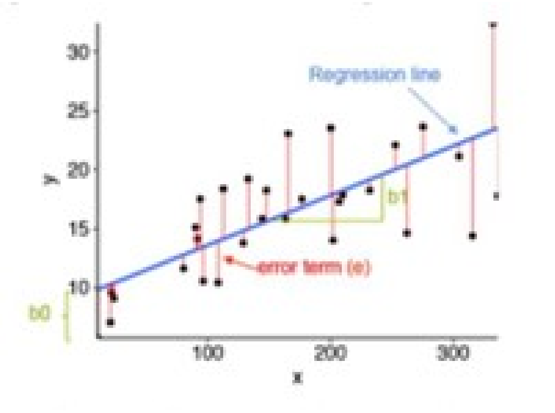

Figure 5: Illustration of linear regression.

3.1.1. *Relationship between actual and predicted values for both models*

From the LR model, the trend lines with fitted equations of $y = 0.55x + 16.31$, do not follow the ideal $y = x$ and have broader dispersion. Results can be considered more accurate if the prediction approaches the diagonal line. The blue dots are much more spread and, therefore exhibit a weak linear relationship between the predicted and actual values. However, this is not the case for the LGBM model. There is a strong linear relationship between the predicted and the actual values. The trend lines with fitted equations of $y = 0.92x + 2.52$, only had a very little dispersion which follows the ideal linear function, $y = x$. Points in the LGBM model are nearer to the diagonal line, therefore the LGBM model performs better, in other words, more accurate than the LR model. This indicates that the values predicted by applying the LGBM model to predict the concrete compressive strength are relatively close to the actual values compared to the LR model. This research also computed the effect of each input variable to see if the eight input variables influence the final compressive strength. The graph in Figure 9 shows the coefficient magnitude of feature importance from the LR model. It can be seen that

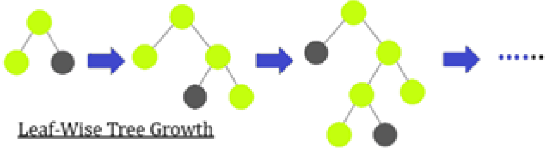

Figure 6: Architecture of LGBM model.

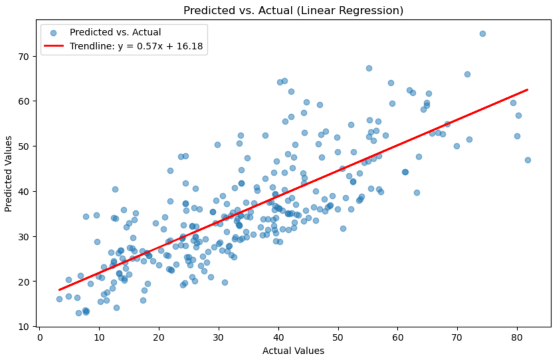

Figure 7: Predicted vs. actual values of the LR model.

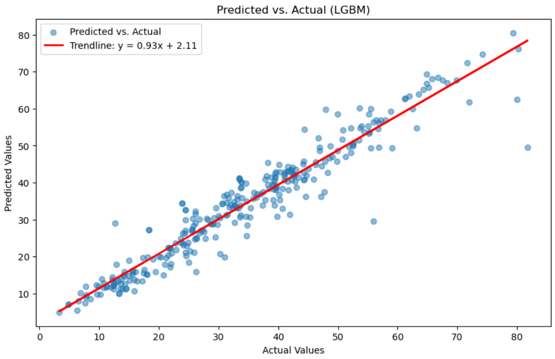

Figure 8: Predicted vs. actual values of the LGBM model.

the only variable exceeding 12 was cement, while the other variables fell under 10. This means that cement is the most contributing variable to the compressive strength, followed by blast furnace slag, age and fly ash. However, water is at the bottom of the graph as the only variable with a negative coefficient of magnitude. This result did not quite follow the assumption mentioned earlier in this paper, where the main and the most important factors of variables should be cement, water, fine aggregate, and coarse aggregate. This may be due to the LR model capturing the linearity of variables in contributing to the compressive strength but does not consider the complexity such as the correlation of variables to each other in contributing to the overall compressive strength. Table 4 shows the coefficient magnitude of feature importance for each variable in descending order. The compressive strength of concrete can be predicted with a high degree of accuracy using the LGBM model developed in this research. To visually represent the impact of these input variables on the LGBM model towards the compressive strength, the graph in Figure 10 is plotted. From this figure, it is shown that the three features that exceeded 2000

Table 4: Coefficient magnitude of feature importance from the LR model

Feature	Importance (coefficient magnitude)
Cement	12.504414
Blast Furnace Slag	9.254417
Age	6.929565
Fly Ash	6.063005
Fine Aggregate	1.797203
Coarse Aggregate	1.702513
Superplasticizer	1.666422
Water	-2.409308

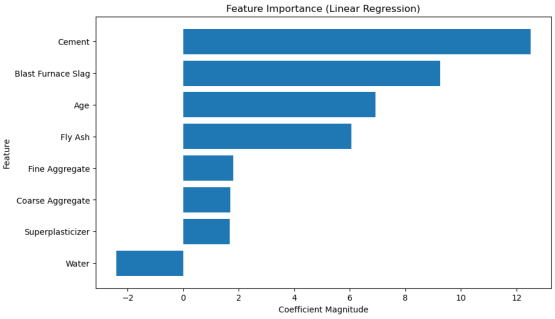

Figure 9: Features importance from the LR model.

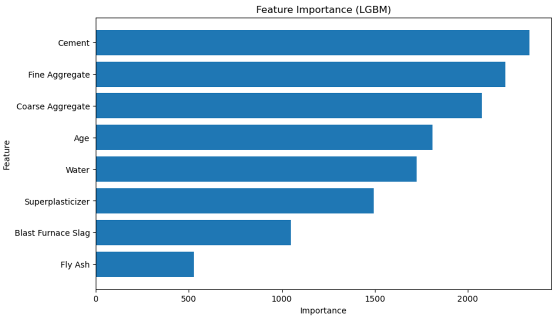

Figure 10: Feature importance from the LGBM model.

Table 5: Importance gain of feature importance from the LGBM model

Feature	Importance (gain)
Cement	2331
Fine Aggregate	2200
Coarse Aggregate	2076
Age	1811
Water	1724
Superplasticizer	1493
Blast Furnace Slag	1050
Fly Ash	527

importance gain where cement, fine aggregate, and coarse aggregate features have a dominant influence on compressive strength, according to the theory discussed. Although we expected the water to fall after the three features, it does not place far below the graph, only to be below the feature age that falls right under coarse aggregate. We can say that LGBM gives us the result aligned with our assumption. Table 4. shows the importance gain of feature importance for each variable in descending order.

4. Conclusion

LR model is a linear based model, where the algorithms only capture the linear relationship between variables X and Y. If we refer to figure 3, the Pearson correlation heatmap shows that cement, age, superplasticiser, and blast furnace slag have an impact on the compressive strength. This is true because the LR model feature importance shows three variables (cement, blast furnace slag, age) out of four from the one shown in the Pearson correlation heatmap. However, just like the Pearson correlation heatmap, the LR model only captures linearity patterns in the datasets, not other complexity factors. LR model is also sensitive to outliers. During EDA, it is important to deal with outliers properly so that they won't affect the model's coefficient and subsequently its prediction. Another factor that must be taken into account is that the LR model does not detect the non-monotonic relationship, in which the variables may not be consistently increasing and decreasing. Lastly, the LR model performs less on a large dataset and with high numbers of features. LGBM is a tree-based model, where its algorithms can capture patterns like complex nonlinear relationships and the interactions between variables. If we refer to the assumption made, we can see that the main ingredients of concrete are cement, water, fine aggregate, and coarse aggregate. We also assume that being the main ingredient will also be the major factor in contributing to the compressive strength. This is proven by the LGBM model where also, out of four mentioned earlier in this paper, three were predicted by the model, which are cement, fine aggregate, and coarse aggregate. LGBM is much more robust in dealing with outliers, which prevents them from affecting its coefficient during prediction. Another reason that the LGBM model behaves more accurately is that it detects the non-monotonic relationship between variables in a dataset. The leaf-wise architecture of the LGBM model allows it to handle larger datasets. In conclusion, the LGBM model performs better than the LR model with the R^2, MAE, MSE and RMSE equal to 0.920, 2.992, 21.613 and 4.649, respectively. LR model, however, has lesser accuracy with R^2 equal to 0.607 while the error values of MAE, MSE and RMSE are equal to 8.207, 106.56, and 10.32, respectively. Therefore, the LGBM model can predict the concrete compressive strength much more precisely compared to the LR model.

Acknowledgements

I would like to thank Praveen Yadav for their support in completing this research.

Authorship contribution

T. Hakimi: Calculation; Analysis; and Writing M. Bhuyan: Supervision and Draft Corrections

Funding

No funding was used in this study.

Conflict of interest

No conflicts of interest.

Declaration

This research has been conducted ethically, reporting of those involved in this article.

Similarity Index

I hereby confirm that there is no similarity index in abstract and conclusion while overall is less than 7% where individual source contribution is 4% or less than it.

Data Availability

Data sharing is not applicable to this article as no data set were generated or analyzed during the current study.

References

[1] T. Han, A. Siddique, K. Khayat, J. Huang, A. Kumar, Construction and Building Materials, 244, 118271(2020). https://doi.org/10.1016/j.conbuildmat.2020.118271

[2] V. V. Degtyarev, M. Z. Naser, Structures, 34, 3391–3403 (2021). https://doi.org/10.1016/j.istruc.2021.09.060

[3] O. R. Abuodeh, J. A. Abdalla, R. A. Hawileh, Applied Soft Computing, 95, 106552 (2020). https://doi.org/10.1016/j.asoc.2020.106552

[4] K. L. Chung, L. Wang, M. Ghannam, M. Guan & J. Luo, Journal of Building Engineering, 35, 101998(2021). https://doi.org/10.1016/j.jobe.2020.101998

[5] Prasad, Concrete ingredients and important aspects, Structural Guide(2023). Available at: https://www.structuralguide.com/concrete-ingredients/.

[6] F. H. Chiew, International Conference on Computer and Drone Applications (IConDA) (2019). https://doi.org/10.1109/iconda47345.2019.9034920

[7] S. H. Kosmatka, W. C. Panarese, B. Kerkhoff, Design and control of concrete mixtures, Vol. 5420, pp. 60077-1083 (2002). Skokie, IL: Portland Cement Association.

[8] C. LeBow, Effect of cement content on concrete performance. University of Arkansas (2018).

[9] M. Saberian, J. Zhang, A. Gajanayake, J. Li, G. Zhang, & M. Boroujeni, Handbook of Sustainable Concrete and Industrial Waste Management, 637–659 (2022). https://doi.org/10.1016/b978-0-12-821730-6.00007-3.

[10] R. Flatt, I. Schober, Understanding the Rheology of Concrete, 144–208 (2012). https://doi.org/10.1533/9780857095282.2.144.

[11] W. Ben Chaabene, M. Flah & M. L. Nehdi, Construction and Building Materials, 260, 119889 (2020). https://doi.org/10.1016/j.conbuildmat.2020.119889

[12] D. C. Feng, Z. T. Liu, X. D. Wang, Y. Chen, J. Q. Chang, D. F. Wei & Z. M. Jiang, Construction and Building Materials, 230, 117000 (2020). https://doi.org/10.1016/j.conbuildmat.2019.117000.

[13] M. R. Kaloop, D. Kumar, P. Samui, J. W. Hu & D. Kim, Construction and Building Materials, 264, 120198 (2020). https://doi.org/10.1016/j.conbuildmat.2020.120198.

[14] A. Ahmad, F. Farooq, P. Niewiadomski, K. Ostrowski, A. Akbar, F. Aslam & R. Alyousef, Materials, 14(4), 794(2021). https://doi.org/10.3390/ma14040794.

[15] S. Chithra, S. R. R. S. Kumar, K. Chinnaraju & F. Alfin Ashmita, Construction and Building Materials, 114, 528–535 (2016). https://doi.org/10.1016/j.conbuildmat.2016.03.214.

[16] S. C. Lee, Engineering Structures, 25(7), 849–857(2003). https://doi.org/10.1016/s0141-0296(03)00004-x.

[17] H. Nguyen, T. Vu, T. P. Vo & , H. T. Thai, Construction and Building Materials, 266, 120950 (2021). https://doi.org/10.1016/j.conbuildmat.2020.120950.

[18] A. Kumar, H. C. Arora, N. R. Kapoor, M. A. Mohammed, K. Kumar, A. Majumdar & O. Thinnukool,Sustainability, 14(4), 2404 (2022). https://doi.org/10.3390/su14042404.

[19] J. Zhang, G. Ma, Y. Huang, J. Sun, F. Aslani & B. Nener, Construction and Building Materials, 210, 713–719 (2019). https://doi.org/10.1016/j.conbuildmat.2019.03.189.

[20] I. C. Yeh, Cement and Concrete Research, 28(12), 1797–1808 (1998). https://doi.org/10.1016/s0008-8846(98)00165-3

Copyright & License