XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Estimation of concrete compressive strength using machine learning

Dimas Betioli Ribeiro[1], Pedro Cesar Rodrigues Alves Abrão[2], Carlos Jose Massucato[2], Vanderley Moacyr John [2], André Carlos Ponce de Leon Ferreira de Carvalho [3]

[1]*Civil Engineering Division, Aeronautics Institute of Technology*
Praça Marechal Eduardo Gomes, 12228-900, São Paulo/São José dos Campos, Brazil
dimasbetioli@gmail.com
[2]*Department of Construction Engineering, Escola Politécnica, University of São Paulo*
Av. Professor Almeida Prado, Travessa 2, 83, 05508-070, São Paulo/São Paulo, Brazil
pedro.abrao@lme.pcc.usp.br, carlos.massucato@massucato.com, vmjohn@lme.pcc.usp.br
[3]*Institute of Mathematical and Computer Sciences, University of São Paulo*
Av. Trabalhador São-carlense 400, 13560-970, São Carlos/São Paulo, Brazil
andre@icmc.usp.br

Abstract. Due to its strength and durability, concrete is one of the primary building materials. These properties are crucial for the safety of concrete-based constructions. Although the final concrete strength can only be known after 28 days of production, building companies often employ concrete in the construction process before this period. Consequently, they may use concrete with strength either much higher or lower than necessary for safety. If the companies can predict the strength of the concrete before 28 days, the construction process can be faster, safer, and cost-effective. Thus, one of the main challenges of the construction industry is making the construction process faster without jeopardizing structural safety. A promising approach to this challenge is to predict concrete compressive strength using the concrete industry data. The main goal of this study is to support construction companies in defining the construction process schedule. For this, it investigates using Machine Learning to predict concrete strength before 28 days. For the Machine Learning experiments, the authors collected a dataset with 236,407 examples and 69 attributes, obtained from four Brazilian concrete industries. These industries do not share standards for data collection. Data curation reduced the dataset to 23 attributes and 56,508 examples. The predictive features were collected from many diverse sources. More than just giving a number for the 28-day-old concrete compressive strength, this study uses random forest quantile regression (RFQR) to generate 95% prediction intervals. By using Random Forest, it is possible to explain how an induced model makes inferences. According to the experimental results, the predictions for the central value (quantile = 0.500) presented a high predictive performance, even when using a small number of the collected features.

Keywords: Concrete Strength, Industry Data, Quantile Regression, Random Forest.

1 Introduction

Because of its strength and durability, concrete is the most widely used building material after water [1]. Despite the need to wait 28 days after the concrete production to know its strength, most companies use it in the construction process before the end of this period. The lack of information regarding the final concrete strength can lead to either safety problems or higher construction costs. An alternative to deal with this problem is the design of tools able to predict the final concrete strength when the concrete is used before the 28th day. The need for such a tool motivated many researchers to produce models to predict concrete compressive strength from datasets.

Although the literature presents mathematical formulations for this task [2], most models use ML approaches after the pioneer study presented in [3]. The authors used a dataset obtained experimentally containing eight components of 240 concrete mixtures and the corresponding mechanical strength after 28 days. Applying a neural network to this data, the error for all examples of the test phase was below 5%. Advancing to recent years, the literature on this research area is vast. We briefly summarize the related research below and refer readers interested in a more detailed bibliographic review to [4].

In [5], the authors compare four algorithms—namely, support vector regression (SVR), multilayer perceptron (MLP), gradient boosting regressor (GBR), and extreme gradient boosting (XGBoost)—to predict concrete com-

pressive and tensile strength. They used 1,133 samples and eight attributes related to mix design variables and age. All models resulted in good accuracy, with a coefficient of determination (R^2) ranging within $[0.90, 0.95]$. The authors in [6] compare six tuned ML algorithms: Decision Tree, Extra Tree, Adaptive Boost (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Method (LGBM), and Laplacian Kernel Ridge Regression (LKRR). The study used a dataset of 1,030 samples with eight attributes (mix design variables and age) and obtained R^2 values between 0.86 and 0.96. They reported using Shapley additive explanations (SHAP) to interpret results, identifying variable importance and input ranges that increase and decrease concrete compressive strength. In another related work, [7], the authors use a tuned Extreme Gradient Boosting Machine (XGBoost) algorithm to predict intervals for the compressive strength of high-performance concrete. The authors also recalibrated the hyperparameters to minimize the overestimation of concrete strength, which reduced R^2 from 0.952 to 0.852. The dataset with 400 examples is available in a public repository. Another recent investigation by [8] compared eight techniques, including linear regression, Ridge, LASSO, Decision Trees, Random Forests, XGBoost, support vector machine (SVM), and artificial neural networks (ANN). Based on 1,030 samples and eight attributes, this study observed generally low R^2 values, with the best performance reaching 0.91. Deep learning approaches have also been explored. For instance, [9] applied deep learning techniques on a dataset of 1,100 samples with ten to twelve attributes, including mix design variables, age, and migration test age. Their model achieved a mean absolute error (MAE) of 3.843. Finally, [10] investigated the impact of various material proportions on concrete strength using Decision Tree (DT), Support Vector Machine (SVM), and AdaBoost Regressor (AR). With a dataset comprising 1,287 samples and eleven attributes—covering ratios for binder, cement, water, and various admixtures—the study reported R^2 values ranging from 0.61 to 0.98.

These studies highlight the advancements in applying machine learning models to predict concrete strength, demonstrating a clear trend towards more sophisticated methods and larger datasets to enhance prediction accuracy. Nevertheless, one thing they all have in common is using high-quality experimental data. On the other hand, the data provided by the concrete industry requires further data treatment operations. To the authors' best knowledge, no works in the literature use this type of data to produce ML models to predict concrete compressive strength. The main object of this work is to fill this gap. Thus, while other works base their results on smaller datasets (usually below 1,300 examples), we had access to a dataset that surpassed 56,000 records after preprocessing. These differences lead to challenges not faced by other authors and solved in this study.

2 Dataset description

Concrete, composed of cement, water, aggregates, and additives, is the most used structural material in civil construction [5]. Its production includes several steps summarized in Fig. 1.



Figure 1. Process of concrete production. After selecting the materials (a) and defining a dosage (b), the concrete is mixed and transported to the construction site (c). There it goes through placement (d), finishing (e) and curing process (f).

Concrete companies accumulate data throughout the process, and some seek to predict concrete strength from this data. Some relevant facts about this data are:

1. The companies require data anonymization, a fact associated with the pre-competitive nature of the project.
2. Even when considering only one company, one finds problems like missing data, mistyped values, text where numbers should appear, non-standardized units, duplicated examples, among others.
3. Variables are not available simultaneously. For example, a highly informative variable, concrete compressive strength after 7 days of age, is unknown before these 7 days.

4. Even after the company has collected all the data, this data is spread into separate sources, and an employee interested in a prediction can only access part of the data.
5. Companies need to understand which variables are worthy of collecting.

The original dataset provided for this study contains 236,407 examples with 69 attributes from four companies. The companies provided data trying to solve issue 1, but some sensitive data remained. This was solved by eliminating some variables, like the region where the concrete was produced. Concerning issue 2, a simple inspection eliminated some variables, including blank and non-informative ones. Further analysis detected severe corruption in the data from one company, which was removed from the dataset. Issue 2 was also addressed using classical preprocessing techniques, like KNNImputer [11]. Tab. 1 presents the variables maintained in the study after these procedures. The first column lists the variable type, the second shows the count of each variable, the third specifies its units, and the fourth provides a description of each variable. The table shows that most variables (14) come from the concrete mix design.

Table 1. Variables Used in this Study

Type	**Number**	**Unity**	**Description**
Concrete mix design	14	kg/m^3	Proportion of materials used in the concrete, recorded on the invoice when the truck arrives at the construction site.
Cement strength	3	MPa	Compressive strength of the cement used to produce the concrete, measured according to Brazilian standards at 3, 7, and 28 days of age.
Concrete strength	1	MPa	Compressive strength of the concrete used to produce the concrete at 7 days of age.
Volume	1	m^3	Quantity of concrete provided to the construction site.
Aggregate density	2	g/cm^3	Density of the two aggregates used in the concrete.
Transportation time	1	minutes	Time for the concrete to be transported from the production centre to the construction site.
Slump test	1	cm	Result of the slump test.

We propose training the ML techniques using five different sets of inputs to solve issues 3 and 4. Tab. 2 presents these sets, where the first column identifies the set, the second presents the number of variables considered for each set, and the third describes these variables:

Table 2. Variable sets used to train and test the ML techniques

Set	**Variables**	**Description**
1	2	Quantity of cement and water in the concrete mix design.
2	5	Variables of set 1 plus cement strengths.
3	6	Variables of set 2 plus concrete strength at 7 days.
4	18	Variables of set 3 plus all remaining data in the concrete mix design.
5	23	All variables.

Input set 1 includes variables CMD_cement and CMD_water. These variables present the proportions of cement and water used in a concrete mix design and are the most basic ones to be provided to predict concrete strength. The input set 2 includes $cement_3d$, $cement_7d$ and $cement_28d$, which are strengths at 3, 7, and 28 days age measured for the cement used in the concrete. The input set 3 adds Fc_7d, the concrete strength at 7 days. For input set 4, the remaining concrete mix design variables added are additives (CMD_Silica, CMD_Plast, CMD_Poli and $CMD_Superplast$), aggregates (CMD_agg_0 and CMD_agg_1), sand (CMD_sand_nat and CMD_sand_art) and complementary values calculated from the previous (CMD_WC, CMD_add, CMD_mortar

and CMD_water_level). The final variables added for input set 5 are the volume of ordered concrete $\mathrm{V}olume$, the density of the used aggregates $density_0$ and $density_1$, the time of transportation $time_transportation$ and the result of the slump test $Slump$. Finally, we solve issue 5 with a sensitivity analysis.

3 Quantile Regression with Random Forest

Quantile Regression (QR) extends ordinary least squares regression by modeling the conditional quantile functions of the response variable. Unlike traditional regression, which focuses on estimating the conditional mean, QR provides a more complete picture of the relationship between predictors and the response by estimating various quantiles (e.g., median, percentiles). This is particularly useful when the data exhibits heteroscedasticity or non-normal error distributions [12].

Random Forest is a powerful ensemble learning technique that constructs many decision trees at training time and outputs the mean prediction of the individual trees. Random forests are known for their high accuracy, robustness to outliers, and ability to handle high-dimensional data [13].

Combining Quantile Regression with Random Forest allows for the estimation of conditional quantiles by leveraging the strengths of both methods. This combination provides a non-parametric approach to quantile regression, making no assumptions about the functional form of the relationship between predictors and the response.

The algorithm for Quantile Regression with Random Forest can be summarized as follows:

1. **Bootstrap Sampling:** Create multiple bootstrap samples (subsets of the training data drawn with replacement).
2. **Tree Building:** For each bootstrap sample:
   (a) Build a decision tree. Select the best split at each node from a random subset of predictors.
   (b) Grow the tree until a stopping criterion is met.
3. **Quantile Estimation:** For a given prediction point, obtain the predictions from all trees in the forest. The quantiles of the distribution of these predictions are then used as estimates of the conditional quantiles.

This approach leverages the diversity of the forest trees to capture the uncertainty in the predictions, allowing for the estimation of any quantile. Some advantages of using random forest with quantile regression are [14]:

- **Non-Parametric:** This technique makes no assumptions about the functional form of the relationship between predictors and the response.
- **Robustness to Outliers:** Random forests are known for their robustness to outliers, which is beneficial in quantile regression as outliers can significantly influence quantile estimates.
- **Handles High-Dimensional Data:** It can handle datasets with many predictors.
- **Captures Non-linear Relationships:** Decision trees allow for the modeling of complex non-linear relationships.
- **Low incidence of overfitting:** A previous version of the code used in this work employed gradient boosting, which led to overfitting, a problem solved by using Random Forest instead.

Quantile regression with random forests can be implemented using the 'RandomForestRegressor' from scikit-learn, combined with a custom quantile estimation function. While scikit-learn does not directly offer quantile regression within its 'RandomForestRegressor', we can leverage the ensemble's predictive power to estimate quantiles post-hoc.

A common approach involves training a standard RandomForestRegressor and using its predictions to estimate the desired quantiles. This approach leverages the power of the Random Forest to capture complex relationships in the data. Subsequent quantile estimation, performed outside the RandomForestRegressor, allows for flexible quantile estimation.

Figure 2 demonstrates a 10-fold cross-validation scenario [15] where the dataset is divided into 10 folds. In each iteration, one fold is designated for validation (red), while the remaining nine folds are utilized for training (grey). This work uses k = 5. The `scikit-learn` library in Python provides convenient functions for implementing k-fold cross-validation. This study uses four metrics to evaluate model performance: the normalized root mean squared error (NRMSE), the normalized mean absolute error (NMAE), the coefficient of determination (R^2), and the mean absolute percentage error (MAPE).

4 Results

Table 3 summarizes the results of the experiments. The models show improved performance around the central value (Quantile = 0.500) as more variables are added to the input set. The most significant improvement occurs when transitioning from set 2 to set 3, where the 7-day concrete strength is introduced as an input variable. This improvement highlights the importance of this variable, as discussed below. While further improvements are

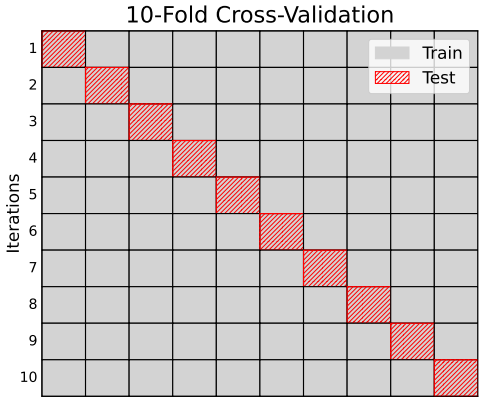


Figure 2. Illustration of 10-Fold Cross-Validation.

observed from set 3 to set 4, the transition from set 4 to set 5 yields no notable changes.

Table 3. Variable sets used to train and test the ML techniques

Input set	Quantil	Mean MAPE	SD MAPE	Mean NMAE	SD NMAE	Mean NRMSE	SD NRMSE	Mean R^2	SD R^2
1	0.025	10.7	0.058	0.050	0.0013	0.068	0.0017	0.68	0.0055
	0.500	10.1	0.058	0.044	0.0011	0.060	0.0014	0.75	0.0038
	0.975	12.1	0.097	0.051	0.0016	0.070	0.0023	0.66	0.0032
2	0.025	12.3	0.113	0.056	0.0017	0.076	0.0023	0.60	0.0077
	0.500	9.3	0.026	0.041	0.0011	0.057	0.0013	0.77	0.0043
	0.975	15.2	0.219	0.061	0.0021	0.084	0.0032	0.51	0.0124
3	0.025	14.9	0.129	0.070	0.0016	0.088	0.0020	0.45	0.0103
	0.500	7.3	0.094	0.033	0.0011	0.046	0.0015	0.85	0.0029
	0.975	19.0	0.205	0.079	0.0026	0.096	0.0031	0.35	0.0135
4	0.025	13.1	0.035	0.060	0.0016	0.075	0.0019	0.61	0.0052
	0.500	6.6	0.018	0.029	0.0009	0.041	0.0012	0.88	0.0021
	0.975	16.4	0.108	0.069	0.0019	0.084	0.0023	0.51	0.0062
5	0.025	13.9	0.111	0.064	0.0015	0.078	0.0019	0.58	0.0073
	0.500	6.4	0.058	0.029	0.0010	0.040	0.0013	0.89	0.0027
	0.975	17.3	0.142	0.073	0.0016	0.087	0.0019	0.47	0.0100

The central prediction (0.500) consistently demonstrates greater accuracy than predictions at the edges (0.025 and 0.975). Considering R^2 alone, the highest accuracy at the edges occurred for input set 1.

Fig. 3 presents the impact of each variable on the model output via SHapley Additive exPlanations (SHAP) [16].

As expected, the most crucial variable is concrete strength at 7 days of age, which is directly proportional to the target variable. Conversely, `CMD_WC` is inversely proportional to the target variable, which was expected since this variable represents the water/cement ratio in the concrete mix design. The second most important variable is first type aggregate density `density_0`, mainly when it presents high values. Cement strength at 28 days of age and transportation time present intermediate importance. `CMD_water` appearing as the second least important variable was a surprise, since it is one of the most basic ones, included in input set 1.

Fig. 4 compares predicted values with real ones, considering input set 5. One can observe a uniform agreement between values. No consistent overestimation or underestimation was observed. The Mean Absolute Error is 3.9 MPa, with many values surpassing 10 MPa. This means this system still needs improvements to be useful for orienting decision-making in real engineering problems.

5 Conclusions

This work presents a methodology for obtaining predictive intervals for concrete compressive strength using variables typically accumulated by the concrete industry throughout concrete production. Other works from the

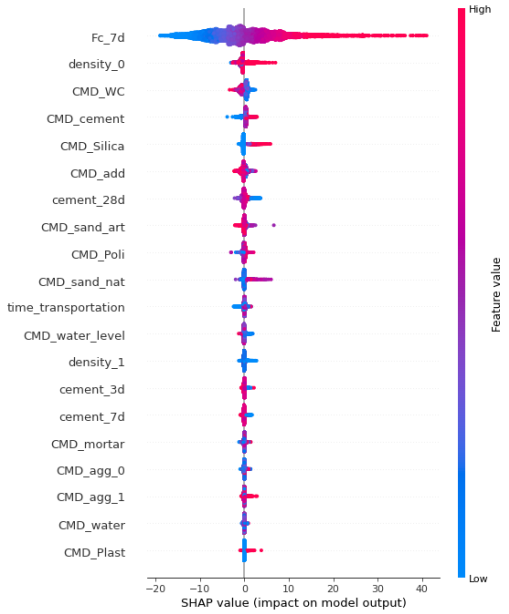


Figure 3. Results of variable importance analysis via SHAP.

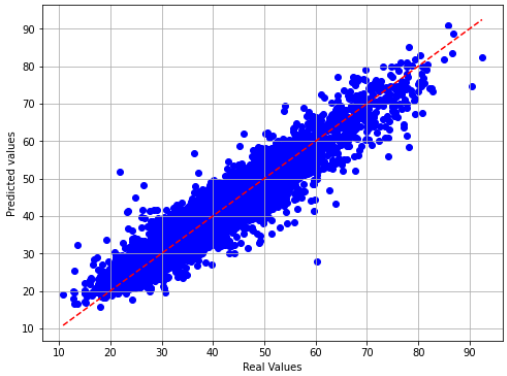


Figure 4. Comparing predicted values to real values.

literature use high-quality experimental data gathered into smaller datasets, which avoids challenges that emerge when dealing with data from the industry. These challenges include missing data and unstandardized variables, among others. To solve these issues, this work proposes creating different sets of inputs that consider real scenarios when users will use the produced ML models. This approach leads to five input sets. Results demonstrate that models reach reasonable accuracy after the set includes the concrete strength at 7-day age, which is the most crucial variable. Concrete mix design data, as well as the time of transportation are also essential. These findings can orient entrepreneurs in deciding which variables to prioritize for data collection throughout concrete production.

Acknowledgment

This project is supported by EMBRAPII CICS POLI USP, and a consortium of nine companies.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] A. G. Ghoniem and L. A. Aboul Nour. Mechanical properties prediction of sandstone concrete with varying compaction levels and silica fume ratios using machine-learning approaches. *Construction and Building Materials*,

vol. 460, pp. 139817, 2025.
[2] S. Popovics. Analysis of concrete strength versus water-cement ratio relationship. *ACI Materials Journal*, vol. 87, 1990.
[3] S. Lai and M. Serra. Concrete strength prediction by means of neural network. *Construction and Building Materials*, vol. 11, n. 2, pp. 93–98, 1997.
[4] R. Kumar, E. Althaqafi, S. G. K. Patro, V. Simic, A. Babbar, D. Pamucar, S. K. Singh, and A. Verma. Machine and deep learning methods for concrete strength prediction: A bibliometric and content analysis review of research trends and future directions. *Applied Soft Computing*, vol. 164, pp. 111956, 2024.
[5] H. Nguyen, T. Vu, T. P. Vo, and H.-T. Thai. Efficient machine learning models for prediction of concrete strengths. *Construction and Building Materials*, vol. 266, pp. 120950, 2021.
[6] I. Ekanayake, D. Meddage, and U. Rathnayake. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, vol. 16, pp. e01059, 2022.
[7] N.-D. Hoang, V.-D. Tran, and X.-L. Tran. Predicting compressive strength of high-performance concrete using hybridization of nature-inspired metaheuristic and gradient boosting machine. *Mathematics*, vol. 12, n. 8, 2024.
[8] S. Elhishi, A. M. Elashry, and S. El-Metwally. Unboxing machine learning models for concrete strength prediction using xai. *Scientific Reports*, vol. 13, n. 1, pp. 19892, 2023.
[9] M. Hosseinzadeh, H. Samadvand, A. Hosseinzadeh, S. S. Mousavi, and M. Dehestani. Concrete strength and durability prediction through deep learning and artificial neural networks. *Frontiers of Structural and Civil Engineering*, vol. 18, n. 10, pp. 1540–1555, 2024.
[10] M. Khan and M. F. Javed. Towards sustainable construction: Machine learning based predictive models for strength and durability characteristics of blended cement concrete. *Materials Today Communications*, vol. 37, pp. 107428, 2023.
[11] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, vol. 17, n. 6, pp. 520–525, 2001.
[12] R. Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.
[13] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.
[14] N. Meinshausen and G. Ridgeway. Quantile regression forests. *Journal of machine learning research*, vol. 7, n. 6, 2006.
[15] S. Ozdemir. *Principles of data science*. Packt Publishing Birmingham, UK, 2016.
[16] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.