Journal of Physics: Conference Series

Prediction of Concrete Properties Using Ensemble Machine Learning Methods

Prediction of Concrete Properties Using Ensemble Machine Learning Methods

D Prayogo, D I Santoso, D Wijaya, T Gunawan and J A Widjaja

Department of Civil Engineering, Petra Christian University, Surabaya - Indonesia
Corresponding author: m21416088@john.petra.ac.id

Abstract. One of the most commonly used materials in civil engineering is concrete; not only is it cheap and strong, but it is also efficient and convenient. The efficiency of concrete is based on the easiness to place and to compact, which is usually known as workability. However, concrete strength and workability works in different ways; hence it is important to divide concrete into two groups: concrete with low workability and concrete with high workability, in order to achieve a more accurate prediction. Since there is a lot of variations of concrete mix designs, the relationship between each mixture is complex and, thus, requires advanced prediction methods in order to find the most accurate relationships between concrete mix proportion and its compression test result.–Recently, many studies have been conducted on applying multiple artificial intelligence (AI) methods in building different complex and challenging prediction models. Thus, this research employs ensemble machine learning methods to precisely forecast compression strength of concrete mix proportion. The accuracy of the proposed method was calculated using two performance measurements. Subsequently, the study has successfully built the prediction model that can accurately map the relationship between concrete mix proportion and compressive strength.
Keywords: ensemble machine learning, slump test, workability

1. Introduction

In civil engineering, concrete compressive strength is one significant criterion when selecting a type of concrete to be used for a specific purpose [1]. Concrete compressive strength is the ability of concrete to carry loads applied on its surface without cracking, and it is measured with concrete compression test. Concrete will be tested after all the other processes have been completed; making concrete cube or cylinder, curing, and waiting for the concrete to reach its maximum strength (usually 28 days). Any small mistake in the course of testing or making the concrete will result in a repeat of the entire procedure, which could be a slow and expensive process [1].

Since any mistake would require waiting for another 28 days, having the ability to determine concrete strength without having to wait for such time will be a great advantage. Machine learning has been proven to be a better technique than other traditional techniques, due to its incredible learning capabilities [2-7]. The models proposed by machine learning are evaluated by two performance measurements, coefficient of correlation (R) and mean absolute error (MAE), to discover the model's accuracy. However, higher water-cement ratio makes the concrete more workable [8], but does result in a lower concrete compression strength. In this research, a classification is done to distinguish between concrete with lower and higher workability, in order to ensure a more precise prediction.

Although machine learning has been proven to be a better technique, it is necessary to choose the best model for every case. This paper, discusses every combination of four models: Linear Regression

Journal of Physics: Conference Series **1625** (2020) 012024 doi:10.1088/1742-6596/1625/1/012024

Analysis (REG), Classification and Regression Tree Analysis (CART), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The purpose of this study is to find the best model for predicting concrete properties, compressive strength, considering its workability.

2. Literature review

2.1. Linear Regression Analysis (REG)

The REG model that is developed from one or multiple relationship between explanatory variables and dependent variables helps to explain the correlation between the variables and the prediction problems [9]. For instance, changes in the dependent variable Y are always affected by variable X. The formula is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots \ + \beta_n X_n + e \tag{1}$$

where *Y* is the dependent variable; $X_1, X_2, \dots, X_n$ is the explanatory variables; β_0 is a constant variable; $\beta_1, \beta_2, \dots, \beta_n$ are regression coefficients; and *e* is the error term.

2.2. Classification and Regression Tree Analysis (CART)

The representation of CART model is a tree-like structure. CART model performs classification from selecting input variables, split points, and minimizing the number of branches through repeated operation to minimize total error. The equations are as follows:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t) \tag{2}$$

$$p(j|t) = \frac{p(j,t)}{p(t)} \tag{3}$$

$$p(j,t) = \frac{p(j)\, N_j(t)}{N_j} \tag{4}$$

$$p(t) = \sum_j p(j|t) \tag{5}$$

where *i* and *j* are categorical variables in each item; $N_j(t)$ is the recorded number of node t in category *j*;

2.3. Chi-Squared Automatic Interaction Detection (CHAID)

The CHAID model is a decision tree model or statistical model proposed by Kass et al. [10]. In order to generate a decision tree, CHAID uses chi-squared test to determine the optimal and significant splits, where continuous predictors are split into categories with approximately equal number of observations. The model will continue to merge and split until no further splits can be performed, where the group result has no differences. The CHAID model also uses various and different methods to measure different data types. For instances, continuous data from F tests are examined, and categorical data are measured through the CHAID [11].

2.4. Artificial Neural Network (ANN)

The ANNs are a group of information-processing models inspired by biological neural networks; ANN model works similarly to that of the human brain, in which neurons are interconnected through synapses [12]. It receives multiple inputs and uses them to make prediction. The processing element has the following characteristics: (1) filtering function to confirm that incomplete data that was inputted to a specific node do not affect the network; and (2) adaptive learning ability to organize the connective weight between nodes. ANNs have multiple input-output systems, and also a basic structure which include an input layer, a hidden layer, and an output layer. The ANNs can be expressed with following equation:

$$\alpha_i = \sigma\left(\sum_j \omega_{ij} o_j\right), \qquad \sigma(x) = \frac{1}{1-e^{-x}} \tag{6}$$

where α_i refers to ANN activities; ω_{ij} is the weight connecting two neurons; o_j is an an output signal of the ANN; x is the activation of i^{th} neuron; and $\sigma(x)$ is the activation function of the ANN that facilitates transformation of inputs into outputs by multiplying the inputs from the processing elements by the corresponding weights.

2.5. Support Vector Machine (SVM)

The SVMs are machine learning models that were first proposed by Vapnik in 1995. They are widely used for classification, forecasting, and regression. An SVM model is used when target variable involves categorical data.

$$f\ (x,\ \omega) = \sum_{j=1}^{n} w_j g_j\ (x) + b \tag{7}$$

where $g_j(x)$ is a set of nonlinear transformations from input space; b is the bias term, w is the weight vector estimated by minimizing the regularized risk function.

2.6. Performance Measurements

The accuracy of the models is tested using the error indicators, namely performance measurements. This study used two statistical method ti compare the actual and prediction values. The two methods are the coefficient of correlation (R) and mean absolute error (MAE). The performance measurements are explained as follows.

The R index shows the linear correlation between two variables, in this study, the prediction and actual values. The minimum and maximum values of R, are -1 and 1, respectively; are calculated as follows:

$$R = \frac{n \sum_{i=1}^{n} y_i \times P_i - \sum_{i=1}^{n} y_i \times \sum_{i=1}^{n} P_i}{\sqrt{n \times \sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2} \times \sqrt{n \times \sum_{i=1}^{n} {P_i}^2 - \left(\sum_{i=1}^{n} P_i\right)^2}} \tag{8}$$

where P_i is the predicted value; y_i is the actual value; and n is the total number of samples.

The MAE is the mean absolute difference between the prediction and actual values; the calculation is as follows:

$$MAE = \frac{\sum_{i=1}^{n} |P_i - y_i|}{n} \tag{9}$$

where P_i is the predicted value; y_i is the actual value; and n is the total number of samples.

3. Experimental method and result

3.1. Data preparation

A hundred and three data records of concrete mix proportion from previous research were used to analyze concrete strength. Every data set has seven input variables and two output variables, such as cement, slag, fly ash, water, superplasticizer (SP), coarse aggregate (CA), fine aggregate (FA), slump test result (slump), and 28-day compressive strength test result (*fc'*), respectively. The data is divided according to their workability level and measured by its slump test result to ensure a more accurate prediction. Concrete mix proportion with lower workability level (slump test result lesser than 12.5 cm) and higher workability level (slump test result equal to or higher than 12.5 cm). The properties and characteristics of the variables of each data set can be seen in Table 1 and Table 2.

Table 1. Statistical description of concrete mix proportion with slump test result lesser than 12.5 cm.

	Cement (kg)	Slag (kg)	Fly Ash (kg)	Water (kg)	SP (kg)	CA (kg)	FA (kg)	Slump (cm)	*fc'* (MPa)
Min	142	0	0	160	6	721	640.6	0	18.52
Max	356	180	239.9	211	19	1002	815	12	58.53
Mean	209.65	114.84	175.98	179.56	9.75	900.20	717.59	2.62	40.24

Table 2. Statistical description of concrete mix proportion with slump test result equal or more than 12.5 cm.

	Cement (kg)	Slag (kg)	Fly Ash (kg)	Water (kg)	SP (kg)	CA (kg)	FA (kg)	Slump (cm)	*fc'* (MPa)
Min	137	0	0	167.3	4.4	708	647.1	13	17.19
Max	374	193	260	240	15	1049.9	902	29	52.65
Mean	235.08	68.53	142.11	201.68	8.23	879.82	745.24	22.00	34.96

3.2. The process of training and model selection
The number of data set for concrete mix proportion with slump test result lesser than 12.5 cm and equal to or higher than 12.5 cm are 21 and 82, respectively. Next, each of them is divided into training set (70%) and test set (30%). Training set is analyzed by the numeric predictor, machine learning methods, to determine the pattern which will be used to predict the result from the same input for test set. Later, the result prediction will be compared with the actual result, using two performance measurements: MAE and R. Fifteen different models comprising single and ensemble methods, were used to get the most accurate prediction.

3.3. Prediction results and comparison
All of the models were compared to the actual strength, and then ranked based on their performances. The models rank and the result from the best models are mentioned in Table 3, Table 4, and Figure 1.

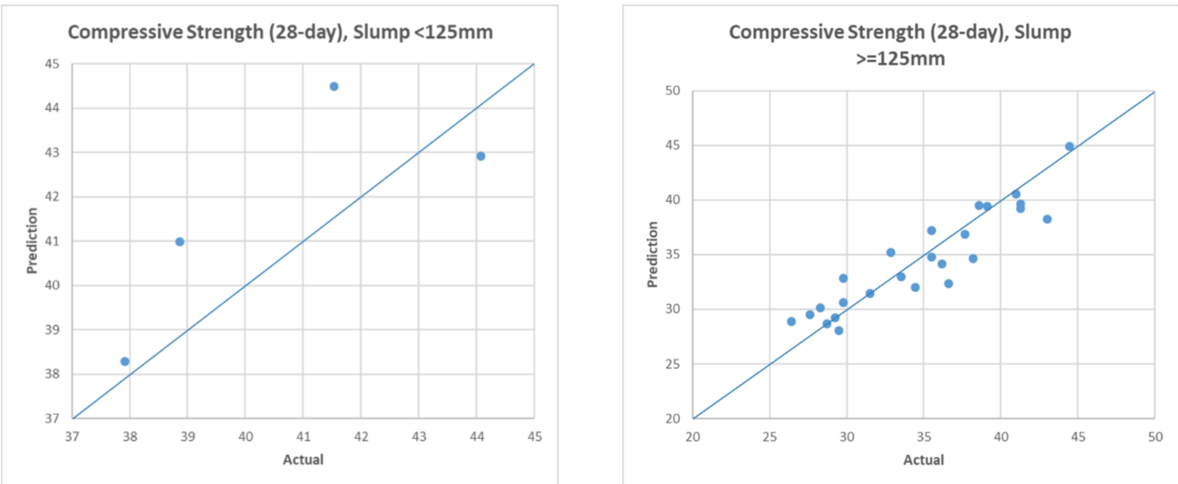
Table 3. Model rank for concrete mix proportion with slump test result lesser than 12.5 cm.

Model	R		MAE		Average Rank
	Value	Rank	Value (MPa)	Rank	
Regression	0.774	2	1.647	1	1.5
Regression + SVM	0.627	4	2.021	3	3.5
Regression + Neural Network + CHAID + SVM	0.441	8	1.968	2	5
Regression + Neural Network	0.777	1	2.589	9	5
Regression + Neural Network + CHAID	0.483	6	2.339	5	5.5
Regression + Neural Network + SVM	0.636	3	2.524	8	5.5
Regression + CHAID + SVM	0.443	7	2.449	7	7
Neural Network + CHAID + SVM	0.274	12	2.102	4	8
Regression + CHAID	0.486	5	3.294	14	9.5
Neural Network + CHAID	0.327	10	2.774	10	10
SVM	-0.467	14	2.394	6	10
CHAID + SVM	0.276	11	2.938	11	11
CHAID	0.331	9	5.118	15	12
Neural Network	-0.284	13	3.261	13	13
Neural Network + SVM	-0.529	15	2.962	12	13.5

Table 4. Model rank for concrete mix proportion with slump test result equal to or higher than 12.5 cm.

Model	R		MAE		Average Rank
	Value	Rank	Value (MPa)	Rank	
Neural Network + SVM	0.922	1	1.623	2	1.5
Regression + Neural Network + CHAID + SVM	0.914	3	1.578	1	2
Regression + Neural Network + SVM	0.912	4	1.649	4	4
Regression + SVM	0.915	2	1.719	6	4
Neural Network + CHAID + SVM	0.907	6	1.642	3	4.5
Regression + CHAID + SVM	0.908	5	1.674	5	5
Regression + Neural Network + CHAID	0.905	7	1.73	7	7
Neural Network + CHAID	0.894	10	1.748	8	9
Neural Network	0.904	8	1.854	10	9
Regression + CHAID	0.891	12	1.795	9	10.5
Regression + Neural Network	0.891	11	1.915	11	11
SVM	0.904	9	2.123	14	11.5
Regression	0.874	13	1.981	12	12.5
CHAID + SVM	0.858	14	2.084	13	13.5
CHAID	0.768	15	2.496	15	15

It is shown on Table 3 that the best model for predicting concrete compressive strength with low workability, slump test result lesser than 12.5 cm, is the Linear Regression (REG) method. However, on Table 4 it is shown that the best model for predicting concrete compressive strength with high workability, slump test result equal to or higher than 12.5 cm, is ensemble Artificial Neural Network (ANN) and Support Vector Machine (SVM) method. Then the prediction of concrete strength is made for every sample in the test set, using the best model for each condition. The comparison between the result from the prediction and the actual value are shown on Figure 1.



(a) (b)

Figure 1. Test results of the regression prediction model for concrete with slump test result (a) lesser than 12.5 cm, and (b) equal to or higher than 12.5 cm.

4. Conclusion

It can be concluded from this research, that one of the most important processes of predicting concrete compressive strength is to divide the concrete based on the workability-slump test result. Since workability and strength works in different ways, dividing the concrete will result in higher accuracy predictions. It can be inferred that the best method for predicting concrete with lower workability, slump test lesser than 12.5 cm, is the Linear Regression (REG) method. On the other hand, the best method of prediction for concrete with higher workability, slump test equal to or higher than 12.5 cm, is ensemble Artificial Neural Network (ANN) and Support Vector Machine (SVM) method.

5. References

[1] Prayogo D 2018 Metaheuristic-based machine learning system for prediction of compressive strength based on Concrete Mixture Properties and Early-Age Strength Test Results *Civil Engineering Dimension* **20**(1) 21-9

[2] Kheder G F, Gabban A M and Abid S M 2003 Mathematical model for the prediction of cement compressive strength at the ages of 7 and 28 days within 24 hours *Materials and Structures* **36**(10) 693

[3] Cheng M Y and Prayogo D 2014 Symbiotic organisms search: a new metaheuristic optimization algorithm *Computers & Structures* **139** 98-112

[4] Cheng M Y, Prayogo D and Wu Y W 2014 Novel genetic algorithm-based evolutionary support vector machine for optimizing high-performance concrete mixture *Journal of Computing in Civil Engineering* **28**(4) 06014003

[5] Cheng M Y, Wibowo D K, Prayogo D and Roy A F V 2015 Predicting productivity loss caused by Change Orders using the Evolutionary Fuzzy Support Vector Machine Inference Model *Journal of Civil Engineering and Management* **21**(7) 881-92

[6] Hoang N D and Pham A –D 2016 Hybrid artificial intelligence approach based on metaheuristic and Machine Learning for Slope Stability Assessment: A Multinational Data Analysis *Expert Systems with Applications* **46** 60-8

[7] Cheng M Y, Prayogo D, Ju Y H, Wu Y W and Sutanto S 2016 Optimizing mixture properties of biodiesel production using Genetic Algorithm-based Evolutionary Support Vector Machine *International Journal of Green Energy* **13**(15) 1599-607

[8] ACI Committee 238 2008 Report on Measurements of workability and rheology of fresh concrete *ACI 238.1R-08* (United States of America: American Concrete Institute) p 45

[9] Skyes A O 1993 An introduction to regression analysis *Coase-Sandor Institute for Law & Economics Working* **20** 1-34

[10] Kass G V 1980 An exploratory technique for investigating large quantities of categorical data *Applied Statistics* **29**(2) 119-27

[11] Chou J -S, Yang K -H and Lin J -Y 2016 Peak shear strength of discrete fiber-reinforced soils computed by machine learning and metaensemble methods *Journal of Computing in Civil Engineering* **30**(4) 04016036

[12] Das S K 2012 Artificial neural networks in geotechnical engineering: modeling and application issues *Metaheuristics Water, Geotechnical Transportation Engineering* (United Kingdom: Elsevier) chapter 10 pp 231-270

Acknowledgments

Authors wishing to acknowledge the present research is financially supported by The Ministry of Research, Technology and Higher Education of the Republic of Indonesia under the Penelitian Dasar Unggulan Perguruan Tinggi (PDUPT) 2019 Research Grant Scheme.