Optimizing Prediction Accuracy of Concrete Mixture Behavior Using Hybrid K-means Clustering and Ensemble Machine Learning

D Wijaya, D Prayogo, D I Santoso, T Gunawan and J A Widjaja
Department of Civil Engineering, Petra Christian University, Surabaya - Indonesia
Corresponding author: m21416216@john.petra.ac.id

Abstract. Concrete is the most used material in infrastructure development, especially in a developing country. The concrete used in project must not only satisfy the desired concrete strength, but also the workability. Additionally, due to different conditions in construction projects, the requirement for workability varies. Workability can be measured using several methods. Previously, traditional trial-and-error of concrete mix design were used to achieve desired slump and flow test value. However, the experiment is often inexpensive, and the obtained results may not be sufficiently accurate. Recently, the potential of the AI method has been gaining increased attention as the new and promising alternative method to predict slump and flow tests, based on historical data. Thus, this study develops an effective hybrid AI-based method to predict slump and flow tests from the given concrete mixture dataset. A total of 103 historical data are used. At the beginning, the samples are separated into two groups using k-means clustering. Each cluster is modelled using the ensemble of six prediction methods, which are REG, CART, GENLIN, CHAID, ANN and SVM. The obtained results show that our proposed method can build the prediction method with a high accuracy, measured by several performance indicators.

Keywords: concrete mixture, K-means clustering, ensemble machine learning methods, optimizing, test, workability

1. Introduction
In the beginning, concrete was only known for its high compressive strength, while in the casting process, it still needed special treatment. Along with the development era, concrete technology is also demanded to keep getting better, not only the desired concrete strength, but also the workability. Concrete mix design is an essential and abstruse topic, which requires extensive knowledge of many expert issues. Obtaining concrete with appropriate strength, and other utility parameters, allows for the reliable use of the structure. The process of concrete hardening and hydration are irreversible [1]. Nowadays, concrete is the most used material in infrastructure development because it possesses high compressive strength, is easy to form and is quite affordable.

During the development era, improvements were also needed for concrete technology; not only regarding the desired concrete strength, but also the workability. Concrete mix design is an essential and abstruse topic, which requires extensive knowledge of many expert issues. Obtaining concrete with appropriate strength, and other utility parameters, allows for the reliable use of the structure. The process of concrete hardening and hydration is irreversible [1]. Many things affect the compressive strength and workability of concrete, the most important being the material used. Concrete is made of cement, coarse
aggregate, fine aggregate, water and other substitution materials. Chemicals are also used to improve the quality of concrete. Superplasticizer is the most used chemicals in concrete.

Based on the compressive strength, concrete is divided into three groups, namely low strength concrete, medium strength concrete and high strength concrete. Low strength concrete is concrete with compressive strength below 20 MPa; medium strength concrete is concrete with compressive strength between 21-40 MPa and high strength concrete is concrete with compressive strength above 41 MPa [2]. Concrete compressive strength refers the compressive strength of concrete when the concrete is 28 days old. In this research, most of the concrete data is of medium quality.

Based on its workability, concrete is divided into two groups, ordinary concrete and self-compacting concrete. Ordinary concrete is conventional concrete that is commonly used in simple buildings. It requires special treatment at the time of casting, which is required conventional compaction or with a compactor or vibrator. Unlike ordinary concrete, self-compacting concrete does not require special treatment during casting. After pouring fresh concrete, the concrete can be solid itself.

The problem that exists in concrete mix design is the number of internal and external factors that affect the concrete compressive strength and workability. Internal factors can be controlled when making concrete, such as the mix design and material, while external factors cannot be controlled, such as external temperatures and aggregate conditions that exist in the field. The second problem is the hardening of concrete, which takes about twenty-eight days to reach its full compressive strength [3-5]. The water needed for full hydration of cement varies from 20% to 25% of its mass, without taking into account the water trapped in the pores [6-7]. Therefore, almost all standard concrete compressive strength tests are done when the sample is 28 days old. The third problem is the relation between workability and concrete compressive strength.

As technology develops, Artificial Intelligence or AI, also participates in statistical development. AI is technology that is used to copy the intelligence of living things to solve a problem. Three methods have been developed, namely Fuzzy Logic (FL), Evolutionary Computing (EC) and Machine Learning (ML). Machine Learning method is an approach in AI that is used to replace or imitate human behavior to automatically solve problems [8]. The methods in AI used in this research are Machine Learning Method and K-means Clustering. K-means Clustering is a method used to divide data sets into two or more groups based on one or more desired variables. In this study, the variable slump test and flow test values from the concrete data set are used. Meanwhile, the Machine Learning Method is used to predict results based on existing training data. It consists of six aspects, namely Linear Regression (REG), Classification and Regression Tree Analysis (CART), Generalized Linear Model (GENLIN), Chi-Squared Automatic Interaction Detection (CHAID), Artificial Neural Network (ANN) and Support Vector Machine (SVM). In this study, the variable to be predicted is the concrete compressive strength. Each method has advantages and disadvantages, so a single method cannot always be the best in predicting every existing data set.

The accuracy of the prediction of the data set is seen from the value of the Coefficient of Correlation (R) and the Mean Absolute Error (MAE). The result of R value shows the precision between the results of the training data and the results of the testing data. Meanwhile, the MAE is the average error testing data when compared with training data. To get a better R value and also minimize the amount of MAE, a new method is used, namely Ensemble Machine Learning, which means combining two or more of the methods available in Machine Learning.

2. Literature Review

2.1. Linear Regression Analysis (REG)
Linear Regression refers to a group of techniques for fitting and studying the straight-line relationship between two variables [9]. Linear regression estimates the regression coefficients β0 and β1 in the equation:

\[ Y = \beta_0 + \beta_1 X_j + \varepsilon_j \]  \hspace{1cm} (1)
where \( X \) is the independent variable, \( Y \) is the dependent variable, \( \beta_0 \) is the \( Y \) intercept, \( \beta_1 \) is the slope, and \( \varepsilon \) is the error.

2.2. Classification and Regression Tree Analysis (CART)

The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree [10]. The 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 algorithm, proposed by a statistician Kass in the late 1970s, is one of the most popular statistically based methods of supervised learning for decision tree development. CHAID algorithm is used for the detection of association between the categorical dependent variable and multiple independent variables, which can be categorical and/or metric. [11]

2.4. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is one of the Machine Learning methods/aspects that imitate human nerves, which is part of the fundamental of the brain. ANN is divided into input layer and output layer [8]. This method uses input layer to make a prediction. The ANNs can be expressed with the following equation:

\[
\alpha_i = \sigma(\sum_j \omega_{ij}o_j), \quad \sigma(x) = \frac{1}{1-e^{-x}} \tag{6}
\]

where \( \alpha_i \) refers to ANN activities; \( \omega_{ij} \) is the weight connecting two neurons; \( o_j \) is 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. K-means Clustering

K-means clustering lies in partitioning the clustering method most frequently used in datamining; the algorithm segregates \( N \) number of documents into \( K \) number of clusters. In this research, we use a \( K \) value of 2 for accuracy of prediction. The goal of K-means is to decrease the summation of square distance among data points and their respective cluster centers [12].

2.6. Performance Measurements

The accuracy of all the method is based on coefficient of correlation (R) value and mean absolute error (MAE). The performance measurements are explained as follows.

2.6.1. Coefficient of Correlation (R). The R value shows the correlation between the prediction value and the actual value. In this research, the value is the 28 days old concrete strength. The range for of R
value is -1 to 1. When the value is near zero, there is no linear relationship. As the correlation gets closer to plus or minus one, the relationship is stronger. A value of one (or negative one) indicates a perfect linear relationship between two variables [9] respectively; R is calculated as follows:

\[
R = \frac{n \sum_{i=1}^{n} y_i \times p_i - \sum_{i=1}^{n} y_i \times \sum_{i=1}^{n} p_i}{\sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2} \times \sqrt{n \sum_{i=1}^{n} p_i^2 - (\sum_{i=1}^{n} p_i)^2}}
\] (8)

where \( p_i \) is the predicted value; \( y_i \) is the actual value; and \( n \) is the total number of samples.

2.6.2. Mean Absolute Error (MAE). The MAE is the mean absolute difference between the prediction and actual values; the calculation is as follows:

\[
MAE = \frac{\sum_{i=1}^{n} |p_i - y_i|}{n}
\] (9)

where \( p_i \) is the predicted value; \( y_i \) is the actual value; and \( n \) is the total number of samples.

3. Methodology

Figure 1. First experiment method.

The first experiment flow is explained in Figure 1. The data set using the excel format is uploaded into clementine software. Then, we set the data that we want to use. Cement, slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, slump test and flow test are set up into input for the program. Concrete strength at 28 days old is set up into output for the program. Next, we divide the data set into two groups, “training data” and “testing data”. From the “training data”, we can get numeric predictor. After that, we use the equation from numeric predictor that we did earlier to predict the “testing data”.

Figure 2. Second experiment method.

The second experiment is slightly different from the first one as shown in Figure 2. This time, the input for the program is cement, slag, fly ash, water, superplasticizer, coarse aggregate and fine aggregate. Slump test and flow tests are used for K-means to divide the data set into two groups, cluster 1 and cluster 2. Each cluster will be done exactly the same as first experiment.

Figure 3. Third experiment method.

The third experiment is exactly the same as the second experiment as shown in Figure 3. However, in order to improve the prediction, we use a new method to test the data, which is the ensemble method. This method uses two or more numeric predictors to test the data. We use eleven types of ensemble methods, namely ANN-CART-REG-CHAID, ANN-CART-REG, ANN-CART-CHAID, ANN-REG-CHAID, CART-REG-CHAID, ANN-CART, ANN-REG, ANN-CHAID, CART-REG, CART-CHAID and REG-CHAID.
4. Experimental Results
The result of the first experiment can be seen on table 1. The accuracy of the prediction from the first experiment is visibly not good enough, with correlation value (R) not more than 0.502 and mean absolute error being more than 6.200 MPa. Since the prediction is not good enough, where the average coefficient of correlation (R) being 0.380 and the average of mean absolute error (MAE) being 6.780 MPa, this method cannot be used to predict the concrete strength. Therefore, we need to improve the method to get a better prediction.

Table 1. Testing result for first experiment.

| Method | Mean Absolute Error (MAE) | Coefficient of Correlation (R) |
|--------|---------------------------|-------------------------------|
| ANN    | 5.785                     | 0.502                         |
| CART   | 7.001                     | 0.302                         |
| REG    | 7.531                     | 0.358                         |
| CHAID  | 6.624                     | 0.337                         |
| SVM    | 6.200                     | 0.421                         |
| GENLIN | 7.531                     | 0.358                         |

The second and the third experiments divide the dataset into two clusters, which can be seen on Table 2 and Table 3. 103 historical data are divided into 76 historical data for the first cluster and 27 historical data for the second cluster. As we can see, the cluster is divided based on the slump and flow test, the other variable is not used.

Table 2. Statistical description of concrete mix proportion for first cluster.

| Cement (kg) | Slag (kg) | Fly Ash (kg) | Water (kg) | SP (kg) | CA (kg) | FA (kg) | Slump (cm) | Flow (cm) |
|-------------|-----------|--------------|------------|---------|---------|---------|------------|-----------|
| Min 137.00  | 0.00      | 0.00         | 168.00     | 4.60    | 708.00  | 647.10  | 13.00      | 40.00     |
| Max 374.00  | 193.00    | 260.00       | 240.00     | 15.00   | 1049.50 | 902.00  | 29.00      | 78.00     |
| Mean 239.45 | 65.27     | 140.17       | 140.17     | 8.26    | 874.78  | 748.85  | 22.54      | 58.66     |

Table 3. Statistical description of concrete mix proportion for second cluster.

| Cement (kg) | Slag (kg) | Fly Ash (kg) | Water (kg) | SP (kg) | CA (kg) | FA (kg) | Slump (cm) | Flow (cm) |
|-------------|-----------|--------------|------------|---------|---------|---------|------------|-----------|
| Min 140.00  | 0.00      | 0.00         | 160.00     | 4.40    | 721.00  | 640.60  | 0          | 18.26     |
| Max 356.00  | 180.00    | 239.90       | 211.00     | 19.00   | 1049.9  | 815.00  | 17         | 58.53     |
| Mean 203.01 | 113.73    | 173.91       | 181.46     | 9.33    | 909.89  | 713.59  | 5.4167     | 38.52     |

However, the results from the second and the third experiments were good enough, because they use k-means clustering. It can be seen in Figure 4 and Table 5 that using an ensemble method of ANN and REG can do the best accuracy prediction, with R value being 0.921 and MAE being 2.030 MPa. While for the second cluster, it can be seen in Figure 5 and Table 7 that by only using the REG method, it can do the best accuracy prediction, with R value at 0.980 and MAE at 1.622 MPa.
| Method                        | Mean Absolute Error (MAE) | Coefficient of Correlation (R) |
|-------------------------------|---------------------------|-------------------------------|
| ANN                           | 2.345                     | 0.921                         |
| CART                          | 2.973                     | 0.796                         |
| REG                           | 2.188                     | 0.907                         |
| CHAID                         | 4.932                     | 0.475                         |
| ANN-CART-REG-CHAID            | 2.552                     | 0.877                         |
| ANN-CART-REG                 | 2.259                     | 0.910                         |
| ANN-CART-CHAID               | 2.856                     | 0.835                         |
| ANN-REG-CHAID                | 2.599                     | 0.877                         |
| CART-REG-CHAID               | 2.845                     | 0.842                         |
| ANN-CART                     | 2.461                     | 0.891                         |
| ANN-REG                      | 2.030                     | 0.921                         |
| ANN-CHAID                    | 3.045                     | 0.806                         |
| CART-REG                     | 2.404                     | 0.887                         |
| CART-CHAID                   | 3.554                     | 0.731                         |
| REG-CHAID                    | 3.029                     | 0.823                         |

**Figure 4.** Graphic for first cluster using ANN-REG.
Table 5. Testing result for second cluster.

| Method                | Mean Absolute Error (MAE) | Coefficient of Correlation (R) |
|-----------------------|---------------------------|-------------------------------|
| ANN                   | 2.98                      | 0.885                         |
| CART                  | 4.748                     | 0.614                         |
| REG                   | 1.622                     | 0.98                          |
| CHAID                 | 5.788                     | 1                             |
| ANN-CART-REG-CHAID    | 3.012                     | 0.94                          |
| ANN-CART-REG          | 2.478                     | 0.94                          |
| ANN-CART-CHAID        | 3.604                     | 0.828                         |
| ANN-REG-CHAID         | 3.35                      | 0.975                         |
| CART-REG-CHAID        | 3.389                     | 0.89                          |
| ANN-CART              | 3.357                     | 0.828                         |
| ANN-REG               | 2.133                     | 0.975                         |
| ANN-CHAID             | 4.214                     | 0.885                         |
| CART-REG              | 2.698                     | 0.89                          |
| CART-CHAID            | 4.557                     | 0.614                         |
| REG-CHAID             | 3.705                     | 0.98                          |

5. Conclusion
In this research, it can be concluded that the dataset needs to be divided into two clusters based on results from the slump test and flow test; this will improve the prediction accuracy (MAE) and correlation between actual and prediction value. The first cluster contains 76 historical data and the second cluster contains 27 historical data. The first cluster is concrete with higher workability based on slump test and flow test. The best method used for the first cluster is an ensemble of 2.4. Artificial Neural Network (ANN) and Linear Regression Analysis (REG) with R 0.921 and MAE 2.030 MPa. However, the best method that is used for the second cluster is the Linear Regression Analysis (REG) with R 0.980 and MAE 1.622 MPa. Since the R is greater than 0.900 and the MAE is less than 2.500 MPa, the method from this research can be used to predict the concrete strength at 28 days old based on material used, the slump test and the flow test.
6. References

[1] Ziolkowski P and Niedostatkiewicz M 2019 Machine Learning Techniques in Concrete Mix Design Materials 12(8) 1256

[2] ACI Committee 2011 Building Code Requirements for Structural Concrete (America: American Standard) ACI 318-11

[3] Marchon D and Flatt R J 2016 Mechanisms of cement hydration. In Science and Technology of Concrete Admixtures Elsevier pp 129–145

[4] Scrivener K, Snellings R and Lothenbach B A 2018 Practical Guide to Microstructural Analysis of Cementitious Materials (USA: CRC Press Boca Raton, FL) ISBN 1498738672

[5] Kurdowski W 2014 Concrete Chemistry (Germany: Springer Science & Business, Berlin) ISBN 9400779453

[6] Young J F, Mindess S and Darwin D 2002 Concrete (USA: Prentice Hall, Upper Saddle River, NJ) ISBN 0130646326

[7] Hover K C 2011 The influence of water on the performance of concrete Constr. Build. Mater. 25 3003–13

[8] Ahmad A 2017 Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning J. T. Ind.

[9] NCSS Statistical Software 2019 Linear Regression and Correlation, chapter 300 Retrieved on 2019 from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Linear_Regression_and_Correlation.pdf

[10] Loh W 2011 Classification and Regression Trees WIREs Dat. Mining and Know. Disc. 1 (1) pp 14-23

[11] Milanović M and Stamenković M 2016 CHAID decision tree: methodological frame and application Eco. Them. 54(4) 563-86

[12] Naeem S and Wumaier A 2018 Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K Intern. J. Comp. App. 975 8887

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 (PDUPUT) 2019 Research Grant Scheme.