Modelling the minislump spread of superplasticized PPC paste using Random forest, Decision tree and Multiple linear regression

M Mrithula¹, Modepu Bhavana², S Janani³, Vaishnavi Narayanan³, Rajesh C.B⁴*, Dhanya Sathyan⁵

¹Department of Electronics and Communications Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
²Department of Civil Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
³Department of Electronics and Communications Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
⁴Assistant Professor(Sr.Grade), Department of Electronics and Communications Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
⁵Associate Professor, Department of Civil Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.

Email: cb_rajesh@cb.amrita.edu
Email: s_dhanya@cb.amrita.edu

Abstract. Workability is one of the key property of concrete which is governed by water cement ratio. In order to improve the workability of concrete without any variations in water cement ratio Superplasticizers(SPs) are added. Cement paste helps us to analyze the property of fresh concrete where the dispersion of cement particle is taken into account. SP’s Cement dispersive properties are governed by dosage and the family. Various dosages and families of SP are considered for estimating workability feature of cement paste which is picked for investigating on rheological properties through Mini slump spread diameter. The prime motive of this analysis includes measuring the workability of different superplasticizers by conducting a minislump test and hence modelling the flow rate of the superplasticized Portland Pozzolona Cement (PPC)paste using the application of random forest(RF),decision tree(DT) and multiple regression algorithms. Testing and training data for a model were 287 unique mixture compositions at a water by cement ratio was 0.37. This mixture was tested experimentally in a laboratory using four types of locally available PPC’s and of SP which can be broadly categorised in to four families. Amount of seven types of SP brands, water content, cement weight were the input parameters for the model and flow rate was the output parameter. The model’s predicted and experimentally measured values of flow speed were compared and the amount of deviation was recorded.

Keywords: Mini slump, Random Forest, Decision Tree, Multiple Regression Algorithms

1. Introduction:

Concrete is used based on the workability, strength, durability and quality. The innovations in the name of chemical admixtures created a bench mark in the construction market by improving the workability at minimal W/C ratio. Chemical Compositions of sulphonate melamine formaldehyde, modified
lignosulphonates, sulphonate naphthalene formaldehyde and copolymers containing sulphonic and carboxyl groups [6]. Most commonly used cement in civil engineering is Portland pozzolona cement (PPC). Pozzolonic materials such as flyash are used in the manufacturing process instead of cement. Which is an added advantage in reducing heat of hydration. The micro filler effect and pozzolonic reaction of pozzolonic material present in PPC results in the increase of durability. PPC is pollution free and eco-friendly material compared to many other available compositions due to the emission of CO₂ in minimum amount during the manufacturing process.

Modelling of mini slump spread using Random kitchen sink (RKS) algorithm was carried out by Dhanya et al (2018)[6]. In that work ingredients quantities were taken as the input parameters. As the family of superplasticizer has great effect on the workability of the mix (Jayasree and Gettu 2002[7], Jayasree and Gettu[9]), superplasticizer from different families were also treated as separate input parameter in this study. In this study it is observed that RKS model is good in predicting the mini slump spread.

Effectiveness of RKS in predicting the fresh and hardened stage properties of self-compacting concrete were studied by Dhanya et al 2018[8]. In this study ingredient quantities were taken as the input parameter and fresh and hardened stage properties were taken as the output parameters. To check the prediction accuracy of the model, RMSE and MAE of the predicted values were calculated [8] and it is found that all values are less than 0.05.

Determining the flow rate of cement paste experimentally every time is a tedious job that requires an experimental set up, it is time consuming and economically not viable. Hence there is a dire need for computing the flow rate of cement paste using soft computing that reduces effort, time and cost for repeated predictions. Machine learning is an evolving field and embodies different ways for novel details[1]. Classification is one of many methods in Machine Learning that helps to group data, whereas the method that attempts to determine the strength and character of the relationship between one dependent variable and multiple independent variables is called regression[2]. So ML regression method estimates the output value using the input samples of the dataset (training dataset). The purpose of the regression method is to minimize the deviation of the predicted outputs from the actual outputs [3].

There are many ML regression methods like multiple regression, decision tree[4], random forest[5]. This work aims at predicting the workability parameters effectively using methodologies that make use of the mean absolute and mean squared error by comparing the experimental and predicted results of models: multiple regression, decision tree[4] and random forest[5]. In the first part of the work laboratory tests have been performed to evaluate the effect of SP on cement paste and in the second part ML regression models were trained to predict flow rate. Prediction capacity of all models were analysed by considering the measured root mean squared error (RMSE), and mean absolute error (MAE).

Python programming language was used to develop the aforementioned models.

2. Approach to modelling and prediction:

In all three algorithms, the train_test_split function has been used to create the appropriate train and test data for the features ("X_train" and "X_test" respectively) and target data ("Y_train" and "Y_test"). The test data is specified to be 20% of the total data.

2.1. Multiple Linear Regression

Regression algorithms are supervised machine learning techniques for predicting continuous numerical values.

\[
y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0
\]
An error variable (disturbance term) is added along with the equation which represents "noise" in the linear relationship between the dependent variable and regressors.

\[ y: \text{response variable} \]

\[ n: \text{number of features} \]

\[ x_n: \text{nth feature (nth predictor variable)} \]

\[ \beta_n: \text{regression coefficient (weight) of the nth feature} \]

\[ \beta_0: \text{y-intercept} \]

This method is used for predicting the outcome of response variable when there are more than two explanatory variables in a dataset. It represents how multiple independent variables are related to one dependent variable ‘y’. The model calculates the best fitted line that minimizes the variances of each of the variables and because it fits a line, it is a linear model. Linear regression focuses on the conditional probability distribution of the response, given the values of the predictors, rather than on joint probability distribution of all of these variables, which is what sets it apart from multivariate analysis.

We have chosen this as one of our algorithms as models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters.

Data that can are modelled using the same can also be represented in the form of matrices.

Where: \( y \) can be equated to the inner product of the transpose of the predictor variable and regression coefficient vectors.

\[
\mathbf{y} = \mathbf{x}^T \mathbf{\beta} + \mathbf{\epsilon}
\]

\[
\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}
\]

\[
\mathbf{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}
\]

In our dataset, the input data has been split into training and validation sets in a 80/20 ratio. The purpose of which is to train the model on the data, where flow rate has been experimentally derived. Test data is then fed into the model to predict flow rate of the same using the predictive model. Linear regression attempts to fit a straight hyper plane that is closest to all data points in the dataset. This algorithm has lower complexity as opposed to other Machine Learning algorithms, the equation of which is fairly easy to understand, assimilate and interpret.
2.2. Decision Tree and Random Forest

A decision tree is a supervised machine learning algorithm which is simply a sequential decision chain that progresses to reach a specific result. The model goes by “if this then that” conditions. The flow of the tree spreads downwards, beginning at the top. It is extremely fast and can work efficiently on large datasets, it is also very easy to visualize. It has two nodes: decision and leaf node. The former is used to make decisions and can branch out, whereas the latter are the output of those decisions and mark the end of branching. True to its name, it starts with a root node and branches on to construct a tree like structure. The ID3 algorithm builds decision trees using a top-down greedy approach. A greedy algorithm, always makes the best choice at any given point.

As easy, decision trees are to understand and interpret, often times, a single tree isn’t enough to provide effective results, which is where the random forest algorithm comes into play. A random forest is but a collection of random decision trees whose results are condensed into one final result. Each node in the decision tree works on a random set of features to calculate the output. It is an example of ensemble learning as it combines the outputs of multiple individual models. What makes it different from decision trees is the randomized feature selection.

For regression, both calculate variance reduction using Mean Square Error in Scikit-learn.

3. Experimental investigation

The flow properties of super plasticized cement paste has been carried about experimentally in laboratories. This study has been included with the locally available four brands of PPC. Standard consistency, initial and final setting times, fineness and specific gravity are important physical properties taken into account as they can affect the flow speed. The cement tests were performed according IS 4031 [10-13]. The obtained test results are given in Table 1. Authors have already performed the modeling of minislump flow values using RKS[6].

| Table 1. Physical properties of the cement |
|-------------------------------|-----|-----|-----|-----|
| Cement | C1  | C2  | C3  | C4  |
| Specific Gravity | 2.85 | 2.80 | 2.80 | 2.72 |
| Fineness (%) | 3.50 | 5.40 | 6.00 | 6.40 |
| Range of Standard Consistency (%) | 36-37 | 35-37 | 36-38 | 36-38 |
| Initial Setting Time (Hours) | 2.67 | 1.83 | 2.37 | 1.25 |
| Final Setting Time (Hours) | 3.67 | 3.08 | 3.23 | 2.58 |

The content of solid particles which are added and the liquid density of the SP were determined as per IS 9103 Annex E[14]. The test results are tabulated in Table 2.

| Table 2: Properties of superplasticizers |
|-------------------------------|-----|-----|-----|
| SP    | Density (g/cc) | Solid Content (%) | pH |
| PCE1  | 1.11           | 36.67           | 6  |
| PCE2  | 1.07           | 29.85           | ≥6 |
| SNF1  | 1.21           | 37.43           | 7-8|
| SNF2  | 1.21           | 37.83           | 7-8|
| LS1   | 1.15           | 30.05           | ≥6 |
| LS2   | 1.17           | 31.61           | ≥6 |
| SMF   | 1.23           | 33.06           | 7  |
3.1 Mini slump test
A mini slump test is a simple and economical method for testing the workability of any kind of fresh cementitious paste. A mini slump mould of a conical shape will be filled with the prepared cement paste. Spread diameter while lifting the mould is noted and considered as the measure of workability.

4. Database preparation (Standardisation)
Since features are numeric we do not need to worry about converting categorical data with techniques. Standardisation rescales the attributes so they have a mean of 0 and standard deviation of 1. It assumes that the distribution is Gaussian alternatively normalisation can be used to rescale between the range of 0 and 1. Scikit-learn's StandardScaler is used in this work.

Table 3: Input variables for training the model (Selected)[6]

| Sl. no | Water   | Cement | SP1  | SP2  | SP3  | SP4  | SP5  | SP6  | SP7  |
|--------|---------|--------|------|------|------|------|------|------|------|
| 1      | 0.154054| 0.0147 | 0.022207| 0    | 0    | 0    | 0    | 0    |
| 2      | 0.153391| 0.0147 | 0.044413| 0    | 0    | 0    | 0    | 0    |
| 3      | 0.155332| 0.0147 | 0     | 0.022228| 0    | 0    | 0    | 0    |
| 4      | 0.154412| 0.0147 | 0     | 0.044456| 0    | 0    | 0    | 0    |
| 5      | 0.160805| 0.0147 | 0     | 0     | 0.044499| 0    | 0    | 0    |
| 6      | 0.160205| 0.0147 | 0     | 0     | 0     | 0.066749| 0    | 0    |
| 7      | 0.158591| 0.0147 | 0     | 0     | 0     | 0     | 0.022936| 0    | 0    |
| 8      | 0.158591| 0.0147 | 0     | 0     | 0     | 0     | 0.045871| 0    | 0    |
| 9      | 0.161371| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0.0445| 0    |
| 10     | 0.160644| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0.066749| 0    |
| 11     | 0.164073| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0     | 0.045895 |
| 12     | 0.163169| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0     | 0.068843 |
| 13     | 0.163735| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0     | 0.045895 |
| 14     | 0.162909| 0.0147 | 0     | 0     | 0     | 0     | 0     | 0     | 0.068843 |

5. Results and Discussions
The mini slump spread’s predicted values using Multiple regression is given in the table 4. Graphical comparison of all the three methods are given in Figure 1.
Table 4: Measured and Predicted Mini slump spread diameter

| Mix number | Measured         | Predicted (Multiple regression analysis) |
|------------|------------------|------------------------------------------|
| 1          | 0.141190442      | 0.15226012                               |
| 2          | 0.142038432      | 0.15227299                               |
| 3          | 0.120861522      | 0.15226016                               |
| 4          | 0.140199365      | 0.15225734                               |
| 5          | 0.108085350      | 0.15257066                               |
| 6          | 0.128762547      | 0.15275454                               |
| 7          | 0.086742623      | 0.15239209                               |
| 8          | 0.120618441      | 0.15254563                               |
| 9          | 0.128001763      | 0.15256142                               |
| 10         | 0.141226117      | 0.15274496                               |
| 11         | 0.124434556      | 0.15259193                               |
| 12         | 0.132730193      | 0.15280669                               |
| 13         | 0.113162034      | 0.1526886                                |
| 14         | 0.135909619      | 0.1529491                                |

The root mean square error (RMSE) and the mean absolute error (MAE) were used to check the prediction accuracy of the model (Refer Table 5). The accuracy of prediction in this model is quite high.

Table 5. Accuracy of prediction

| Sl.no | Machine Learning Algorithms     | Root Mean Square Error | Mean Absolute Error |
|-------|---------------------------------|------------------------|--------------------|
| 1     | Multiple Linear Regression      | 0.0193                 | 0.0151             |
| 2     | Random Forest                   | 0.0160                 | 0.0116             |
| 3     | Decision Tree                   | 0.0162                 | 0.0119             |
6. Conclusion

The flow spread determined through mini slump test method is a good indication for the relative fluidity of superplasticized paste.

RMSE values of prediction of Multiple linear regression, Random forest and Decision Tree are 0.0193, 0.0160, 0.0162 respectively and MAE values of prediction of Multiple regression, Random forest and Decision Tree are 0.0151, 0.0116, 0.0119 respectively. Prediction accuracy of Random forest is found to be better than Multiple linear regression and Decision Tree.

7. References

[1] M. Awad and R. Khanna, Efficient Learning Machines: e-stories, Concepts, and Applications for Engineers and System Designers, Apress, New York, NY, USA. (2015)
[2] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, Boca Raton, FL, USA. (2013)
[3] J. R. Quinlan, Induction of decision trees, Machine Learning, vol. 1, no. 1, pp. 81–106. (1986)
[4] E. Alpaydin, Introduction to Machine Learning, MIT Press, Cambridge, MA, USA. (2009)
[5] A. Liaw and M. Wiener, Classification and regression by randomForest, R News, vol. 2, pp. 18–22. (2002)
[6] Dhanya Sathyan, K B Anand, Chinnu Jose and Aravind N R, Modelling the mini slump spread of superplasticized PPC paste using RLS with the application of random kitchen sink, IOP conference series, Volume 310. DOI:10.1088/1757-899X/310/1/012035. (2018)
[7] Jayasree C and Ravindra Gettu, Experimental study of the flow behaviour of superplasticized cement paste, Materials and Structures, 4 pp. 1581–1593. (2002)
[8] Dhanya Sathyan, K B Anand, Aravind J Prakash and Pemjith B, Modeling of fresh and hardened stage properties of self-compacting concrete using random kitchen sink algorithm. International journal of concrete structures and materials, Springer, DOI 10.1186/s40069-018-0246-(2018).
[9] Jayasree C, Gettu R Influence of mixing method on the fluidity of superplasticized cement paste. Fifth Asian Symposium on Polymers in Concrete, Chennai, India, pp 665–670. (2006)
[10] IS 4031, Part 5, Method for Determination of Initial and Final Setting Time of Cement, Bureau of Indian standards, New Delhi, India. (2005)
[11] IS: 4031, Part 1, Indian Standard Specification for Method of Physical Tests for Hydraulic Cement, Determination of Fineness by Dry Sieving, Bureau of Indian standards, New Delhi. (2005)
[12] IS: 4031, Part 11, Indian Standard Specification for Method of Physical Tests for Hydraulic Cement, Determination of Density, Bureau of Indian standards, New Delhi. (2005)
[13] IS: 4031, Part 4, Indian Standard Specification for Method of Physical Tests for Hydraulic Cement, Determination of Consistency of Standard Cement Paste, Bureau of Indian standards, New Delhi. (2005)
[14] IS 9103, Concrete Admixture-Specification, Bureau of Indian standards, New Delhi, India. (2004).