PREDICTION OF CEMENT STRENGTH USING MACHINE LEARNING

Parth Patel

Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Charotar University of Science and Technology (CHARUSAT) ,
Anand, India.
parth.642001@gmail.com

Poojan Vadaliya

Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Charotar University of Science and Technology (CHARUSAT) ,
Anand, India
poojanvadaliya@gmail.com

Parth Singh

Department of Computer Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Faculty of Technology and Engineering (FTE),
Charotar University of Science and Technology (CHARUSAT),
Anand, India.
parthsingh.dcs@charusat.ac.in

Parth Goel

Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Charotar University of Science and Technology (CHARUSAT) ,
Anand, India.
er.parthgoel@gmail.com

Amit Ganatra

Department of Computer Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT) ,
Anand, India.
amitganatra.ce@charusat.ac.in

Abstract- Compressive strength is one of the most crucial components of concrete design. Time and money can be saved by accurately measuring the factors influencing the compressive strength of concrete. As the compressive strength mainly comprise of nine factors which are cement, blast furnace slag, fly ash, water, superplasticizer, coarse and fine aggregate and its age. So here we have examined the compressive strength affecting by the quantity of the described factors by implementing various model to search for the most well-suited model for different type of material in different condition where it would describe which set of quantity would give the best comprehensive strength. This all can be achieved with the help of machine learning concepts. There are various machine learning models which when trained in a correct manner can provide the solution. Linear Regression and Random Forest are the two models that are substantially used by us. The dataset used is taken from Kaggle, named concrete_data.csv. The paper gives insights of the data used with the help of visualization and also proposes a flow and procedure and proposed technique to achieve better accuracy of the result.

Keywords- Machine Learning, Compressive Strength, Concrete, High Strength Concrete (HSC), Linear Regression, Random Forest, Prediction, k-Means

I. INTRODUCTION

Concrete is widely used around the world due to its cost-effective, monolithic, modular, and long-lasting benefits. High-strength concrete (HSC), which is defined by the compressive strength of more than 40 MPa [1], was pioneered in the field of cementitious materials in the late 1950s and early 1960s. Nowadays, HSC has been widely used in large-span bridges, high-rise buildings, and piers due to its uniform high density, low impermeability, and high durability. [2]

The quality of concrete is determined by its compressive strength. A conventional crushing test on a concrete cylinder is usually used to determine strength. Engineers must construct compact concrete cylinders using various combinations of raw materials and test them for strength differences when each raw material is changed. The recommended wait time for testing the cylinder is 28 days to ensure correct results. To achieve accurate results, it is recommended that you wait 28 days before testing the cylinder. The preparation and testing of several prototypes takes a long time and a lot of effort. Furthermore, this system is vulnerable to human error, and even a minor blunder might result in a significant increase in wait time.

Digital simulations, in which we may supply information to the computer about what we know and the computer attempts alternative combinations to estimate compressive strength, is one technique to reduce wait time and reduce the amount of options to test. We can decrease the number of options we can test physically and the amount of time we spend experimenting in this way. This paper focusses on providing a framework through which the above-mentioned process can be fast tracked.

The main contribution/goal of the paper :

- To reduce the time taken to test the strength
- To reduce the human error, which may be brutal if ignored.
- To make the process cost effective.
- To give results with more accuracy
- To provide result by facts and not by experiments.

II. RELATED WORK

There are various notable works dome by the scholar researchers in this field earlier. One of the notable work titled as “Compressive Strength Prediction of High-Strength Concrete Using Long Short-Term Memory and Machine Learning Algorithms”,[7] by Honggen Chen, Xin Li, Yanqi Wu, Le Zuo, Mengjie Lu and Yisong Zhou which was published in MDPI. They proposed a long short-term memory (LSTM) model was proposed to predict the HSC compressive strength using 324 data sets with five input independent variables, namely water, cement, fine aggregate, coarse aggregate, and superplasticizer. The prediction results were compared with those of the conventional support vector regression (SVR) model using four metrics, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (R2). [7]

III. DATA SET DESCRIPTION

We used Kaggle dataset for the research.

The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory. Data is in raw form (not scaled).

Number of instances (observations): 1030

Number of Attributes: 9

Attribute breakdown: 8 quantitative input variables, and 1 quantitative output variable

Missing Attribute Values: None [3]

Name	Data Type	Measurement	Description
Cement (component 1)	quantitative	kg in a m3 mixture	Input Variable
Blast Furnace Slag (component 2)	quantitative	kg in a m3 mixture	Input Variable--Blast furnace slag is a nonmetallic coproduct produced in the process. It consists primarily of silicates, aluminosilicates, and calcium-alumina-silicates
Fly Ash (component 3)	quantitative	kg in a m3 mixture	Input Variable- it is a coal combustion product that is composed of the particulates (fine particles of burned fuel) that are driven out of coal-fired boilers together with the flue gases.
Water (component 4)	quantitative	kg in a m3 mixture	Input Variable
Superplasticizer (component 5)	quantitative	kg in a m3 mixture	Input Variable--Superplasticizers (SP's), also known as high range water reducers, are additives used in making high strength concrete. Their addition to concrete or mortar allows the reduction of the water to cement ratio without negatively affecting the workability of the mixture, and enables the production of self-consolidating concrete and high performance concrete
Coarse Aggregate (component 6)	quantitative	kg in a m3 mixture	Input Variable--construction aggregate, or simply "aggregate", is a broad category of coarse to medium grained particulate material used in construction, including sand, gravel, crushed stone, slag, recycled concrete and geosynthetic aggregates
Fine Aggregate (component 7)	quantitative	kg in a m3 mixture	Input Variable—Similar to coarse aggregate, the constitution is much finer.
Age	quantitative	Day (1~365)	Input Variable
Concrete compressive strength	quantitative	MPa	Output Variable

[3]

IV. METHODOLOGY

We will need past data on compressive strength of the concrete. After that, we will need to clean the data, as we need to check for irrelevant data and inconsistencies in it. We will be performing a short Exploratory Data Analysis to get the insights from the data. Then, we will use the data for prediction but as we know, there are plenty of models available for the task. So, we found a solution. Before feeding the data to machine learning algorithm, we will do clustering of data. The clustering is needed because there is a mathematical observation that when the entire data is segregated into different clusters and then different models are applied to individual clusters, then accuracy is far more increased than applying different models to the entire dataset.

Proposed Flow:

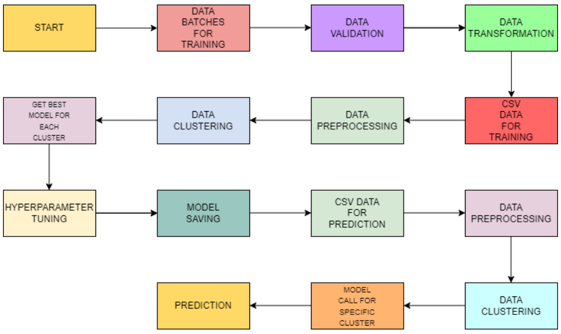

Figure 1 Proposed flow of the application

Model Training Approach

Data Pre-Processing

a) Check for null values in the columns. If present, impute the null values using the KNN imputer

b) transform the features using log transformation

c) Scale the training and test data separately

Clustering - KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using "KneeLocator" function. The idea behind clustering is to implement different algorithms. To train data across several clusters. The Kmeans model is trained on preprocessed data and then saved for later prediction usage.

Following the creation of clusters, we select the best model for each cluster. "Random forest Regressor" and "Linear Regression" are the two algorithms we're utilising. Both methods are passed with the optimal parameters produced by GridSearch for each cluster. The Rsquared scores for both models are calculated, and the model with the highest score is chosen. Each cluster's model is chosen in the same way. For prediction, all of the models for each cluster are kept.

V. PREDICTION

1) Data Export from Db – The data in the stored database is exported as a CSV file to be used for prediction.

2) Data Preprocessing

a) Check for null values in the columns. If present, impute the null values using the KNN imputer

b) transform the features using log transformation

c) Scale the training and test data separately

3) Clustering – Kmeans model created during training is loaded, and clusters for the preprocessed prediction data is predicted.

4) Prediction – Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster.

5) Once the prediction is made for all the clusters, the predictions along with the original names before label encoder are saved in a CSV file at a given location and the location is returned to the client.

VI.NEED AND ROLE OF CLUSTERING

For clustering the data we will be using the k-Mean clustering. K-Means algorithm based on dividing [4] [5] is a kind of cluster algorithm, and it is proposed by J.B.MacQueen. This unsupervised approach is commonly used in data mining and pattern identification. The square-error and error criterion are the underpinnings of this technique, which aims to minimise cluster performance index. This method tries to discover K divisions that match a given requirement in order to find the best result. First, select some dots to represent the initial cluster focal points (typically, we select the first K sample dots of income to represent the initial cluster focal point); second, gather the remaining sample dots to their focal points in accordance with the criterion of minimum distance; third, obtain the initial classification; and finally, if the classification is unreasonable, we will modify it (calculate each cluster focal point again), iterate repeatedly until we obtain the final classification.

K-Means algorithm based on dividing is a kind of cluster algorithm, and has advantages of briefness, efficiency and celerity. [6]

To find the value of k, we used a library named kneed. By this, we can programmatically find the value of the elbow which in turn will give the appropriate cluster number. Once, we find out which model is best for which cluster, we will be saving that model.

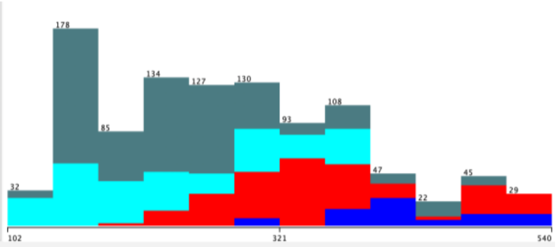

Figure 3 Cement Data analysis Cluster

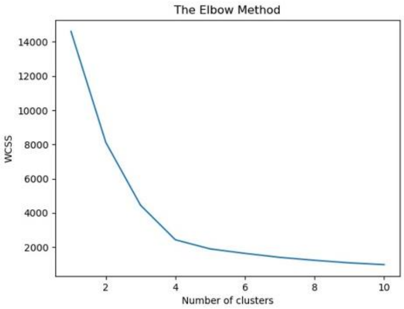

Figure 4 Elbow Graph

As we have selected two models, linear regression and random forest, to decide which model to undertake by which cluster, we will do hyperparameter tuning. This will be done by a function named get_best_models(). This method will compare the results of hyperparameter tuning for both the models and then will determine the best out of them.

VII.MODELS USED:

Linear regression

Regression is a technique for predicting a goal value using independent predictors. This method is primarily used for forecasting and determining cause and effect relationships among variables. The number of independent variables and the form of relationship between the independent and dependent variables are the main differences in regression techniques. Simple linear regression is a type of regression analysis in which there are only one independent variable and the independent(y) and dependent(x) variables have a linear relationship. The linear equation shown below can be used to model the line for linear regression is $y = x * a1 + a0$ Now, the Cost function assists us in determining the best possible values for a0 and a1 in order to obtain the best possible fit line for the data points. For getting the best values of a0 and a1, we transform this search problem into a minimization problem in which we want to reduce the difference between the expected and actual values.

Random forest

The decision tree is used in the random forest, which is made up of a large number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes become the prediction of our model. Any of the individual constituent models will outperform a large number of relatively uncorrelated models (trees) working as a committee. The secret is the low correlation between models. Uncorrelated models can generate ensemble predictions that are more reliable than any of the individual predictions, similar to how low-correlation investments (like stocks and Figure 1 A single tree working in a random forest algorithm. Bonds) come together to create a portfolio that is greater than the sum of its parts. The trees shield each other from their individual mistakes, which results in this wonderful effect.

IX.MODEL SELECTION PARAMETER:

The Coefficient of Determination

To select the model for to evaluate the data of a particular cluster here we have decided to target the model which would be having higher “r2 score” value. Also known as coefficient of determination it would help us to determine the score according to the amount of variance in predictions done on the various data which we could also consider that the score to be calculated by obtaining the difference between the dataset and the values which are predicted by the model. We can observe the r2 score here by the below equations.

$$SS_{res} = \Sigma(y_i - f_i)$$

$$SS_{tot} = \Sigma(y_i - \overline{y})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R2 has several definitions, only some of which are comparable. Simple linear regression using r2 instead of R2 is one of these situations. r2 is just the square of the sample correlation coefficient (i.e. r) between the observed result and the observed predicted value given only one intercept. R2 is the square of the multiple correlation coefficient when more regressors are added. In both cases, the coefficient of determination is usually between 0 and 1.

X.OUTPUT SCENARIOS OF THE MODEL

Models	MAE	RMSE	Correlation Coefficient
Linear Regression	8.01	10.04	0.81
Random Forest	3.56	4.83	0.96

REFERENCES

[1] Henry, G.R. ACI Defines High-Performance Concrete. Concr. Int. 1999, 21, 56–57

[2] Mbessa, M.; Péra, J. Durability of high-strength concrete in ammonium sulfate solution. Cem. Concr. Res. 2001, 31, 1227–1231. [CrossRef]

[3] I-Cheng Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998).

NOTE: Reuse of this database is unlimited with retention of copyright notice for

Prof. I-Cheng Yeh and the following published paper:

I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

[4]JAIN A K, DUBES R C. Algorithms for clustering data[M].New Jersey:Prentice-Hall,1988.

[5]ZhangYufang etc. A kind of improved K-means algorithm [J]. Computer Application,p31ˉ33, 2003, (8).

[6] Youguo Li, Haiyan Wu,

A Clustering Method Based on K-Means Algorithm, Physics Procedia, Volume 25, 2012, Pages 1104-1109, ISSN 1875-3892,

https://doi.org/10.1016/j.phpro.2012.03.206.

[7] H. Chen, X. Li, Y. Wu, L. Zuo, M. Lu, and Y. Zhou, “Compressive Strength Prediction of High-Strength Concrete Using Long Short-Term Memory and Machine Learning Algorithms,” Buildings, vol. 12, no. 3, p. 302, Mar. 2022, doi: 10.3390/buildings12030302. [Online]. Available: http://dx.doi.org/10.3390/buildings12030302