Predicting Concrete Compressive Strength using Machine Learning

Osama Ibrahim
Department of Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
osamani2@illinois.edu

Praneeth Shivashankarappa
Department of Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
ps104@illinois.edu

Kazi Ishat Mushfiq
Department of Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
mushfiq2@illinois.edu

Georg Bauer
Department of Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
georgb2@illinois.edu

Abstract—This project investigates the prediction of concrete compressive strength using machine learning models. We use the Concrete Compressive Strength dataset from the UCI Machine Learning Repository, which contains 1030 experimental data points with eight input variables (cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age) and one output variable (compressive strength in MPa). Our study compares linear regression, decision tree regression, and neural networks to determine the most effective approach for modeling the nonlinear relationships that influence the development of concrete strength.

Index Terms—Concrete Compressive Strength, Machine Learning, Civil Engineering, Regression Models

I. DATASET DESCRIPTION

The Concrete Compressive Strength dataset originates from laboratory experiments conducted by Prof. I-Cheng Yeh, who donated it to the UCI Machine Learning Repository in 2007 to support research on high-performance concrete [1], [2]. The dataset records the quantities of concrete mix ingredients and the curing age, together with the corresponding compressive strength as the output variable. In total, it includes 1030 samples in a CSV file, with each row representing one concrete mix. All ingredient quantities are reported per cubic meter of concrete, and the compressive strength is expressed as a continuous variable in megapascals (MPa), representing the material's capacity to withstand compressive loads.

The dataset has no missing values and is widely recognized in the civil engineering community as a benchmark for modeling concrete strength. For this project, we will obtain the dataset directly from the UCI Machine Learning Repository [2], ensuring a reliable source. The variables included are listed below.

Variables included in the dataset:
- Cement (kg/m³)
- Blast Furnace Slag (kg/m³)
- Fly Ash (kg/m³)
- Water (kg/m³)
- Superplasticizer (kg/m³)
- Coarse Aggregate (kg/m³)
- Fine Aggregate (kg/m³)
- Age (days, from 1–365)

Target variable:
- Concrete Compressive Strength (MPa)

II. SAMPLE DATA PREVIEW

Table I provides representative entries from the Concrete Compressive Strength dataset, illustrating the input features and the output variable. Each row corresponds to one concrete mix, with ingredient quantities and age shown alongside the measured compressive strength in MPa.

TABLE I
REPRESENTATIVE ENTRIES FROM THE CONCRETE COMPRESSIVE STRENGTH DATASET

Row	Cement	Slag	Fly-Ash	Water	SuperP.	CA	FA	Age	Strength
1	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
2	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
4	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
5	198.6	132.4	0.0	192.0	0.0	978.0	825.0	360	44.30
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
1028	148.5	139.4	108.6	192.7	6.1	892.4	780.0	28	23.70
1029	159.1	186.7	0.0	175.6	11.3	989.6	788.9	28	32.77
1030	260.9	100.5	78.3	200.6	8.6	864.5	761.5	28	32.40

III. Project Proposal

A. Planned Approach

In this project, the Concrete Compressive Strength dataset will be analyzed from a data science perspective. Techniques learned in CEE 492 will be applied to perform exploratory data analysis, develop machine learning models, and investigate how each ingredient in the dataset influences the resulting compressive strength. This begins with ensuring the dataset is properly structured for analysis, followed by looking at the distributions of the variables and their relationships. We also plan to explore whether Principal Component Analysis provides useful insights into correlations among the input variables.

After the exploratory phase, we will develop predictive models for compressive strength. Multiple linear regression will serve as the baseline, against which we will compare more advanced approaches such as decision tree models and neural networks. Finally, we will test the trained models on the dataset and check how well they can predict compressive strength, using common accuracy measures.

B. Relevance

Compressive strength is one of the most important properties of a concrete mix, and being able to estimate it accurately is essential in civil engineering, as it can help engineers decide whether a given mixture will meet the required strength for a project. Laboratory testing is the standard approach, but it can be time-consuming and resource-intensive. Empirical equations from design codes can be used as an alternative, but they often lack accuracy and flexibility across different mix proportions and curing conditions, since concrete compressive strength is a highly nonlinear function of both age and ingredient proportions. This creates a need for faster, more reliable methods of prediction.

This project addresses that need by using machine learning to test how well different models can predict compressive strength. A good model can give faster and more consistent estimates of concrete strength for different mix proportions and curing ages, making it a useful tool in engineering practice. The project will also show which ingredients are most important and how curing age affects strength, helping to better understand the factors that influence concrete performance.

C. Deliverables

The deliverables include an exploratory analysis of the dataset, the development and testing of different models to predict compressive strength, a comparison of their performance, and a discussion of the results to show how mix design and curing age influence concrete strength.

IV. Deliverable 2: Exploratory Data Analysis and Predictive Modeling Plan

A. 1. Exploratory Data Analysis

a) *1.1 Dataset Overview:* The Concrete Compressive Strength dataset contains 1030 samples with 8 input features and one output variable. As shown in Table (1).

b) *1.2 Summary Statistics:*

| Feature	Min	Max	Mean	Median	Std-Dev
Cement (kg/m3)	102.0	540.0	281.17	272.9	104.51
Blast Furnace Slag (kg/m3)	0.0	359.4	73.9	22.0	86.28
Fly Ash (kg/m3)	0.0	200.1	54.19	0.0	64.0
Water (kg/m3)	121.75	247.0	181.57	185.0	21.36
Superplasticizer (kg/m3)	0.0	32.2	6.2	6.35	5.97
Coarse Aggregate (kg/m3)	801.0	1145.0	972.92	968.0	77.75
Fine Aggregate (kg/m3)	594.0	992.6	773.58	779.51	80.18
Age (day)	1	365	45.66	28.0	63.17

c) *1.3 Data Visualizations:*
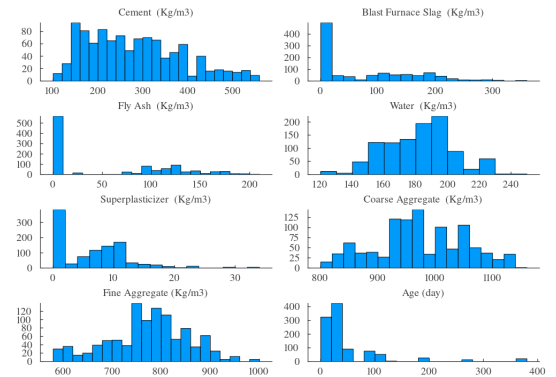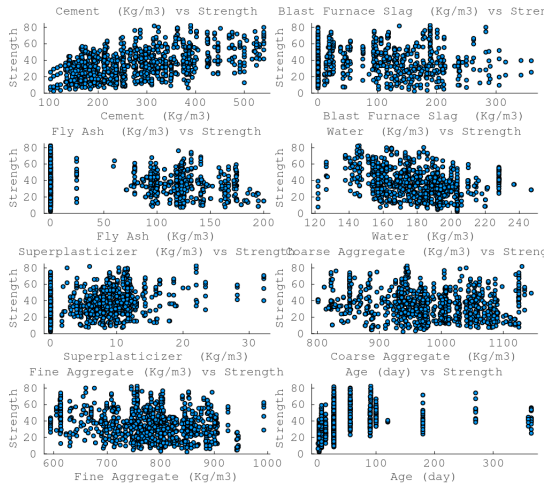

Fig. 1. Histograms of input features

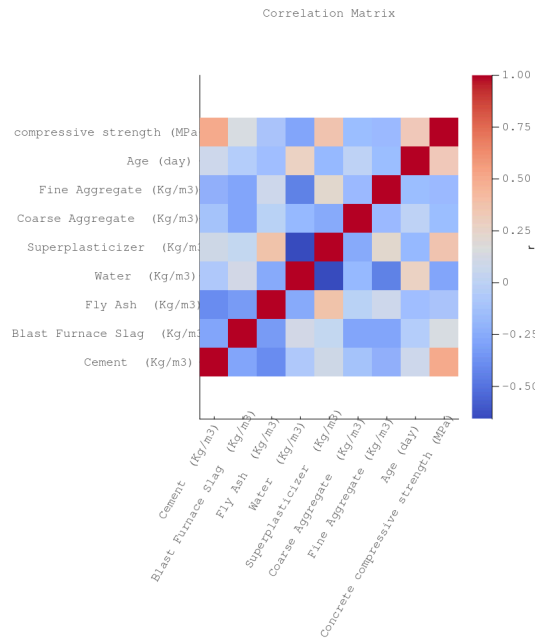Fig. 2. Scatter plots of input features vs compressive strength
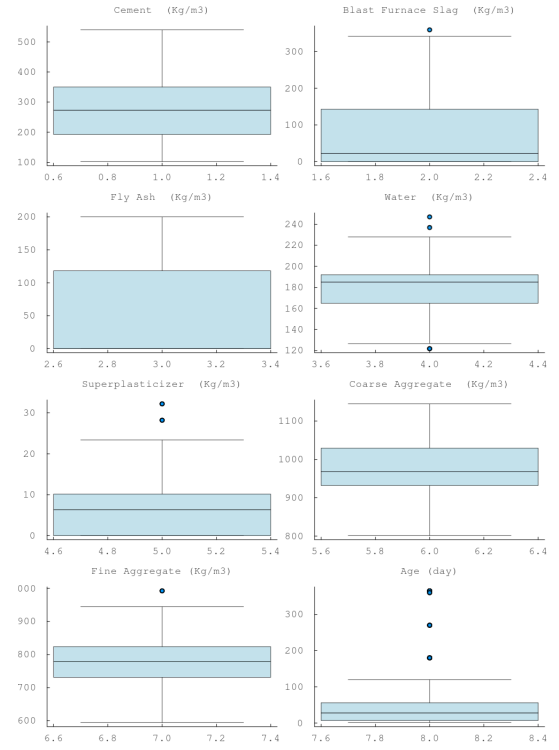

Fig. 3. Correlation heatmap
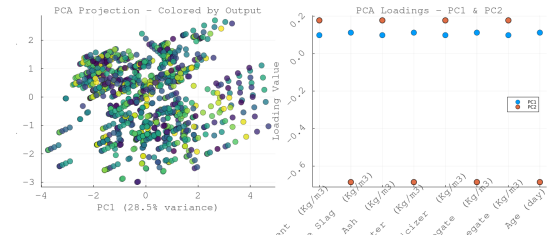

Fig. 4. Boxplots of input features

Fig. 5. PCA plot of input features

B. 2. Predictive Modeling Plan

- Baseline: Linear Regression
- Advanced: Decision Tree, Random Forest, Neural Networks
- Evaluation Metrics: R², RMSE, MAE
- Cross-validation: k-fold (k=5 or 10)
- Feature Importance Analysis

REFERENCES

[1] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, 2007, doi: 10.1016/S0008-8846(98)00165-3.
[2] D. Dua and C. Graff, "UCI Machine Learning Repository: Concrete Compressive Strength Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength