Problem Statement: Concrete Strength Prediction

Objective

To predict the concrete strength using the data available in file **"concrete.csv"**. Apply feature engineering and model tuning to obtain 85% to 95% accuracy.

Resources Available

The data for this project is available in file https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/. The same has been shared along with the course content.

Steps and Tasks:

- **Exploratory Data Quality Report Reflecting the Following:**

1. Univariate analysis –data types and description of the independent attributes which should include (name, range of values observed, central values (mean and median), standard deviation and quartiles, analysis of the body of distributions/ tails, missing values, outliers, duplicates(10 Marks)
2. Bi-variate analysis between the predictor variables and between the predictor variables and target column. Comment on your findings in terms of their relationship and degree of relation if any. Visualize the analysis using boxplots and pair plots, histograms, or density curves. (10 marks)
3. Feature Engineering techniques(10 marks)
    1. Identify opportunities (if any) to extract new features from existing features, drop a feature(if required) Hint: Feature Extraction, for example, consider a dataset with two features *length* and *breadth*. From this, we can extract a new feature *Area* which would be *length * breadth*.
    2. Get the data model ready and do a train test split.

3. Decide on the complexity of the model, should it be a simple linear model in terms of parameters or would a quadratic or higher degree.

- **Creating the Model and Tuning It:**

4. Algorithms that you think will be suitable for this project. Use Kfold Cross-Validation to evaluate model performance. Use appropriate metrics and make a DataFrame to compare models w.r.t their metrics. (at least 3 algorithms, one bagging and one boosting based algorithms have to be there). (15 marks)
5. Techniques employed to squeeze that extra performance out of the model without making it overfit. Use Grid Search or Random Search on any of the two models used above. Make a DataFrame to compare models after hyperparameter tuning and their metrics as above. (15 marks)

Attribute Information:

Given are the variable name, variable type, the measurement unit, and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

	Name	Data Type	Measurement	Description
1	Cement (cement)	quantitative	kg in a m3 mixture	Input Variable
2	Blast Furnace Slag (slag)	quantitative	kg in a m3 mixture	Input Variable
3	Fly Ash (ash)	quantitative	kg in a m3 mixture	Input Variable
4	Water(water)	quantitative	kg in a m3 mixture	Input Variable
5	Superplasticizer (superplastic)	quantitative	kg in a m3 mixture	Input Variable

6	Coarse Aggregate (coarseagg)	quantitative	kg in a m3 mixture	Input Variable
7	Fine Aggregate (fineagg)	quantitative	kg in a m3 mixture	Input Variable
8	Age(age)	quantitative	Day (1~365)	Input Variable
9	Concrete compressive strength(strength)	quantitative	MPa	Output Variable