# Project Problem Statement - FMST

# 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 a score above 85%.

## 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.

## 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 |

## Steps and Tasks:

1. Exploratory Data Quality Report Reflecting the Following:
   1. 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. (**2 Marks**)
   2. Missing values analysis, Outlier detection, Duplicates check, Comment about if the zero values mean the null value here. (**4 Marks**)
   3. Using various plots, write the insights/observations about it. (**6 Marks**)
2. Bi-variate analysis between the predictor variables and also 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 and write the insights/business understanding of the same. (**8 marks**)
3. Feature Engineering techniques
   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. (**3 Marks**)
   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. (**5 Marks**)
4. Creating the Model and Tuning It:
   1. Algorithms that you think will be suitable for this project.
   2. 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 and try to achieve 90% accuracy on testing set). (**7 Marks**)
   3. 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. (**8 Marks**)
5. Write the significance of the R2 score from the business perspective. Explain the effect of an increase/decrease in the R2 score on the business model. (**4 Marks**)
6. Write down the inference about the data analysis and write the recommendations which can be beneficial to the business based on the whole analysis. Also, write the reason for your recommendation. (**8 Marks**)
7. The marks for the score will be according to the following: **testing score 90 -95 (5 marks), 85 - 90 (4 marks), 80 - 85( 3 marks), <80 (2 marks)**