**Regression**
**Project**: To Predict the Strength of Concrete

**Project Description:**

This project focuses on building a predictive model to estimate the **compressive strength of concrete**, a key indicator of its quality and durability. The dataset contains various quantitative attributes representing the composition of concrete mixtures, including the quantities of **cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate**, and the **age of the concrete** in days.

The main objective is to explore how these components influence the **compressive strength** of concrete (measured in MPa) and to develop a regression model that can accurately predict this value. Through **exploratory data analysis (EDA)**, **feature engineering**, and **model training**, this project aims to provide insights into optimal concrete mix design for achieving high strength.

**Main Libraries to be Used:**

- Pandas for data manipulation, aggregation
- Matplotlib and Seaborn for visualisation and behaviour with respect to the target variable
- NumPy for computationally efficient operations
- Scikit Learn for model training, model optimization, and metrics calculation

**Project should include:**

1. **Problem Statement**
2. **Import libraries**
3. **Load dataset**
4. **Data cleaning**
    - Handle missing values
    - Convert data types
    - Remove duplicates
5. **Exploratory Data Analysis (EDA)**
    - Visualize distributions
        i. Univariate analysis
        ii. Bivariate analysis
        iii. Multivariate analysis
    - Correlation analysis
    - Feature-target relationships
6. **Outlier treatment**
    - Boxplot
7. **Check distributions & apply transformations (if needed)**

    - Skewness/Kurtosis
    - Log Transformation, sqrt
8. **Feature engineering**
    - Create new features
    - One-hot encoding (for categorical)
9. **Split data into train/test sets**
10. **Train Linear Regression model**
11. **Feature Scaling**
12. **Prediction using the algorithm**
13. **Visualise the Predicted and Actual**
14. **Print the difference between Actual and Predicted**
15. **Evaluate model performance**
    - MAE - Mean Absolute Error
    - MSE - Mean Squared Error
    - RMSE - Root Mean Squared Error
    - $R^2$ Score - R squared
    - Adjusted $R^2$
16. **Perform the same steps for**
    - Decision tree Regressor
    - Random Forest Regressor
    - Support Vector Machine
    - K Nearest Neighbor
17. **Perform the Cross Validation using Cross_val_score for all the algorithms**
18. **Print the final Conclusion**

**Link to Dataset:** https://github.com/rahulinchal/SPPU

**Data Description:**

1. **Cement** (component 1) -- quantitative -- kg in a m3 mixture -- Input Variable
2. **Blast Furnace Slag** (component 2) -- quantitative -- kg in a m3 mixture -- Input Variable
3. **Fly Ash** (component 3) -- quantitative -- kg in a m3 mixture -- Input Variable
4. **Water** (component 4) -- quantitative -- kg in a m3 mixture -- Input Variable
5. **Superplasticizer** (component 5) -- quantitative -- kg in a m3 mixture -- Input Variable
6. **Coarse Aggregate** (component 6) -- quantitative -- kg in a m3 mixture -- Input Variable
7. **Fine Aggregate** (component 7) -- quantitative -- kg in a m3 mixture -- Input Variable
8. **Age** -- quantitative -- Day (1~365) -- Input Variable
9. **Concrete compressive strength** -- quantitative -- MPa(megapascals) -- Output Variable

## Project Architecture: