**Project name:**

**Revisiting a Concrete Strength regression**

**Names:  
احمد عماد عبد السلام**

**احمد جابر الدغيدي**

**الاء ايمن القاضي**

**About the dataset:**

**Concrete Compressive Strength Data Set**

* Concrete is the most important material in civil engineering.
* The concrete compressive strength is a highly nonlinear function of age and ingredients.

**Data Set Information:**

**Number of instances** 1030  
**Number of Attributes** 9  
**Attribute breakdown** 8 quantitative input variables, and 1 quantitative output variable  
**Missing Attribute Values** None

**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**

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

# 1. Data Visualization and Analysis

# EDA (univariate)

# The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it.

# By using histogram to display the distribution of each variable

## Inferences from EDA

* The data is very much skewed. We need to transform it to nearly normal distribution.

# EDA (bivariate)

# Bivariate analysis uses to show the relationship between the two variables.

# 

## Here, we can find the correlation between the variables

* Compressive Strength increases when Cement, slag, superplasticizer, and age increases (positive correlation).
* Compressive Strength increases when less Water, flyash, coarseaggregate, and fineaggregate is used (negative correlation).

**Removing Outliers**

Before training the model on the training data, we remove the outliers from the training data for the model to be accurate. So we used the boxplot to visualize the data for each feature and remove all outliers for each feature. According to IQR (Inter Quartile Range), any thing not between the lower (minimum) or the upper (maximum) will be considered outlier, according to the figure below:

Chart

Description automatically generated

# 2. Choosing models and training without feature selection

* Split the data set in train and test sets with a proportion of 20% for test set
* Apply the model
* Calculate MAE and R^2 performance metric for the above model

**Using linear regression:**

Mean Absolute Error: 5.42 (the closer to 0 the better)

R2 Score: 0.81

**Using Lasso regression:**

Mean Absolute Error : 10.00 (the closer to 0 the better)

R2 Score: 0.47

**Using Ridge Regression:**

Mean Absolute Error : 5.31 (the closer to 0 the better)

R2 Score: 0.82

**Using Random Forest Regression:**

Mean Absolute Error : 4.41 (the closer to 0 the better)

R2 Score: 0.877

**3. Feature Selection**

**Heatmap**

Treemap chart

Description automatically generatedIn this section we visualized each feature correlation with the other features in the form of a heatmap as in the figure below:

However, it appears from the last row that the highest correlation between the csMPa (output variable) is highest with the following features: cement, superplasticizer, age and water.

**PCA – Principal Component Analysis**

In this section, we performed principal component analysis on the concrete data. The training set features were 8 features which are cement, slag, flyash, watersuperplasticizer, coarseaggregate, fineaggregate and age of the concrete and our output variable (Dependent Variable) is the strength of this concrete. However, when PCA was done, the attained variance is shown in the figure below:Chart

Description automatically generatedThe explained variance for the first 4 principal components didn't even reach 90%. More accurately, 86-79%. However, for the case of our training set and training data for this specific model, it appears that PCA in this regression model is not as important so we will not perform pca for feature selection in this model. As the number of features is small, so we will train the model on all the given features.

**RFE and SelectPercentile**

At last using two methods for feature selection which are RFE (recursive feature elimination) and select percentile for selecting the highest best features to train the model on. For example, in our model we chose 5 features to be best. In case of RFE, it chose cement, slag, water, superplasticizer and the age features. For the selectpercentile it chose all features except slag and flyash.

**4. Choosing model after feature selection**

**In the linear regression model:**

Mean Absolute Error: 4.41 (the closer to 0 the better)

R2 Score: 0.877

We notice that the accuracy of the model (R2 score) increased and the MAE decreased after applying some feature selection & dimension reduction methods to the model.

**In Lasso Regression:**

Mean Absolute Error: 10.00 (the closer to 0 the better)

R2 Score: 0.47

**In Ridge Regression:**

Mean Absolute Error: 5.45 (the closer to 0 the better)

R2 Score: 0.81

**In Random Forest Regression:**

Mean Absolute Error: 4.59 (the closer to 0 the better)

R2 Score: 0.867

After comparing between the results of these models, we found that Linear Regression Model is the best fit model.