Project 12 Concrete strength regression

Dataset

About data:

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

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 (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 -- Output Variable

Procedure

1. Import Data
2. Check dataset size
3. Find and treat missing values (If any)
4. Check column types and describe which columns are numerical, or categorical
5. Perform Univariate analysis
    1. Calculate mean, median, std dev, and quartiles of numerical data
    2. Plot histogram for categorical variables (if any)
    3. Check the distribution of numerical variables and comment on it
6. Perform Bivariate analysis
    1. Plot pair plots
    2. Perform Chi-square analysis to check whether there is a relationship between
        - age and csMPa
    3. Calculate Pearson correlation, and plot their heatmap
7. One hot encode categorical variables (if any)
8. Split into train and test set
9. Scale the variables
10. Train multiple models like Linear regression, Decision Tree, Random Forest, SVR, etc.
11. Check their performance, and comment on which is the best model

Compulsory

1. Use grid search CV to tune the hyperparameter of the best model
2. Train a polynomial regression model with degree 2, and 3 and compare it's performance with other models