Homework 2 – Regression

Problem 1

The dataset "US_CO2.csv" contains the total mass of CO_2 emitted by the United States (in tons/year, column 2) between 1800 and 2014 (year is given is column 1). By training a regression model (method of your choice), predict what the future amount of U.S. CO_2 emissions in 2020, 2050, and 2100. Note that there is no right or wrong answer, but your overall approach needs to make sense and be based on the principles discussed in the lecture and tutorials (e.g., ensuring that your model is not under- or overfitted, etc.). Please briefly explain your approach in your report.

Bonus: Conduct some feature engineering, selection, and regularization to improve the model. Identify the problems in your model and explain why you use these techniques to solve them in your report.

Problem 2

The dataset "Concrete_Data.csv" gathers some measured compressive strength data of various concretes with different formulations. It contains the following information:

1) **Input features:**
   a. Column 1: Fraction of cement (mass %)
   b. Column 2: Fraction of slag (mass %)
   c. Column 3: Fraction of fly ash (mass %)
   d. Column 4: Fraction of coarse aggregates (mass %)
   e. Column 5: Fraction of fine aggregates (mass %)
   f. Column 6: Water-to-cementitious ratio
   g. Column 7: Dosage of superplasticizer admixture
   h. Column 8: Age of the concrete (in days after initial mixture)
2) **Output label:**
   a. Column 9: Measured concrete compressive strength (MPa)

Notes:

- Columns 1-to-3 corresponds to the mass% of cementitious materials (cement, slag, and fly ash) that react with water to form a cement paste.
- Columns 1-to-5 corresponds to the mass% of the solid constituents of concrete (cement, slag, fly ash, coarse aggregates, and fine aggregates). The sum of these fractions is 100%.
- Column 6 corresponds to the ratio of the mass of water to the mass of cementitious materials (cement + slag + fly ash). This dimensionless ratio is commonly used to characterize how much water is added to concrete.
- Column 7 corresponds to the dosage of superplasticizer admixture, expressed as the mass of admixture per mass of cementitious material.
- Column 8 is the age of the concrete (number of days after being mixed with water) when its compressive strength is measured.
- Column 9 is the measured compressive strength of concrete (in MPa), measured by standardized testing method.

Using 10-fold cross-validation, train a regression model (method of your choice) that predicts the compressive strength of concrete. Try to optimize the model so as to minimize the average RMSE of the validation set. Show the intermediate steps of hyperparameter selection. Explain your approach in the report and provide the average RMSE of the training and validation sets (obtained by 10-fold cross-validation) achieved by your model.