Problem set 6

Here, we are working with the famous Diabetes dataset. Our goal is to figure out how to best predict disease progression (y) with the different features in X. Then, let’s visualize that relationship. In order to get a feel for the different variables, check out this webpage for a description of the different meanings of the variables: <https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset>

1. Load the following dataset using the following code:

import pandas as pd

from sklearn.datasets import load\_diabetes

diabetes = load\_diabetes()

X = pd.DataFrame(diabetes.data,columns=diabetes.feature\_names)

y = diabetes.target

From the above code, you should have X as a dataframe with different features. The variable y is your target array.

1. Import other necessary libraries next
2. Create a function that:
   1. Takes X and y as in the input
   2. Split the X and y into train and test using train\_test\_split from sklearn
   3. Fit a linear regression model to the training X and training y data
   4. Get the predicted y data for X test from the model
   5. Output (at least) the model fit (i.e., R^2 score = coefficient of determination)
3. Show that this function works (i.e., try it with some of the data)
4. Then, using the function, find the best single feature in X that best predicts y
5. Using object oriented programming, make a figure that has two subplots. On the left, put the training data. On the right, the testing data. Include the best fitting model line in both plots and include the R^2 value in both plots for the different fits.
6. BONUS (1 point) = Do Steps 5 and 6 inside of a for loop that goes through all of the features. You may have to output more information from your function in Step 3 to plot the right feature!