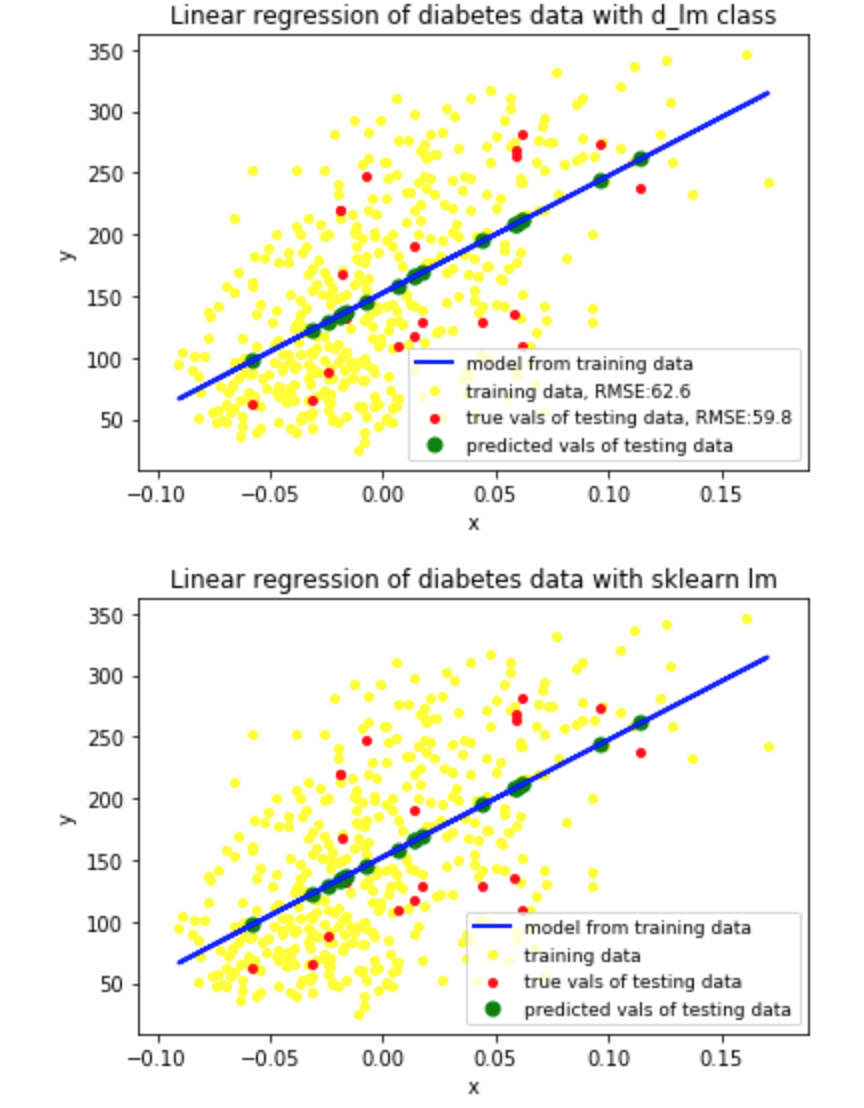
Aaron Yerke, Assignment 1, for Machine Learning Fall 2019

The figures show the results of my linear regression class (top) and the Sci-kit learn’s (sklearn) linear regression (bottom) on the sklearn diabetes dataset. The figures are identical other than the legend, as my linear model class had Root Mean Square Error (RMSE) scores calculated. The blue line shows the calculated model and the yellow and red dots show the training and testing data, respectively. Predicted values are shown along the model in green.

This assignment required us to segregate the data into a training set and a testing dataset. The testing dataset was 20 random points (~4.5 percent) from the total data-set and the training dataset was the rest. Based on my human-innate pattern recognition skills, the model seems to fit both datasets in a similar and fairly recognizable way. In this iteration of the training/testing datasets, the linear regression model seemed to fit the testing data a little better than the training data, based on the RMSE scores (shown in top figure).

However, it other iterations, where a different random 20 points were chosen as the testing dataset, the RMSE of the training set was much better than the testing dataset. This means that the model overfit the training data. This highlights the importance of finding the balance when splitting your data into testing and training sets. Perhaps using a larger percentage of the data as testing data may help with this issue.