Problem 1

Find out how well can you predict the O_3 and NO_2 using the method suggested by the manufacturer. To do this, learn the best linear model that uses just the 4 voltage values to predict O3 and NO2 values. Remember that for this part, you cannot use non-linear models, nor can you use temp, humidity, time stamp as features. However, you can use different loss functions e.g. least squares loss, absolute loss, -insensitive loss as well as different regularizers e.g., ridge, lasso, etc. If you are trying out support vector regression for this part, remember to use the linear kernel. **Describe the method that gave 4 you the best-performing linear model (in terms of MAE on training data), and write down what mean absolute error (MAE) does your model give on the training set.** (10 marks)

Solution: The method which gave us the best answer was using Ridge Regression. The best MAE we got on the training set is 6.08217. The hyperparameters we used were alpha = 6 and tolerance = 10^{-6}.

Problem 2

Chances are that you may not get a very satisfactory result using just a linear model and just the voltage features. Thus, in this next part, develop a learning method that is free to **use temp, humidity, time stamp in addition to the voltage features** to predict the O_3 and NO_2 values. You are also free to use non-linear models e.g. decision trees, kernels, nearest-neighbors, deep-nets, etc. **Describe the method you found to work best giving all details of training strategy e.g. choice of loss function and tuning of hyperparameters.**

Note that you may or may not find the time stamp as a useful feature since some of these pollutants are known to have a diurnal cycle e.g. Ozone is known to have high values during the daytime when sunlight is abundant and low values during night time due to darkness.

Solution First of all, we modified the 'Time' column and changed its value to int $100 \times HH + MM$ and then took the absolute value of difference of this from 1300 because the value of O_3 was highest at about 13:00. Then we scaled the train dataframe and using its mean and variance values, we scaled the test dataset. Then we trained two separate tree models with the same hyperparameters for O_3 and NO_2 prediction. The hyperparameters we used were as follows:-

$ccp\_alpha = 0.01$
$max\_depth = 30$
$min\_samples\_split = 10$
$min\_samples\_leaf = 10$
$min\_impurity\_decrease = 0.001$