Third homework of TOML

# Introduction

The third project consists in the calibration of an air pollution sersor in a network. Precisely, this sensor measure these data:

* The amount of ozone in Kohm;
* The temperature in Celsius;
* The percentage of relative humidity;
* The amount of nitrogen dioxide;
* The amount of nitrogen monoxide;
* The amount of sulfur dioxide;
* The amount of PM10.

Moreover we have the measurements relative to the O3 concentration in ugr/m^3, this is the Reference station, the value we have to predict for the calibration.

To perform our purpose, we used different machine learning models in order to see which model fits better.

# What we used

To perform this project, we used the followings Python libraries:

* Pandas: a library for data manipulation;
* Scikit learn: a library containing many machine learning models in order to perform our project.

# Data Analysis

Before calibrating the sensor, we analyse the data to verify if there is correlation between the reference station value and the measured ones.

## Reading the data

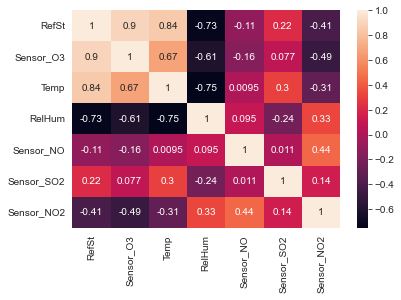
To read the data, we implemented the functions sensorData() and prepareData(), they allow us to read each csv files, make a unique dataframe and add some columns in order to simplify the plotting.

Immagine che contiene testo

Descrizione generata automaticamente Immagine che contiene testo

Descrizione generata automaticamente

## The correlation matrix

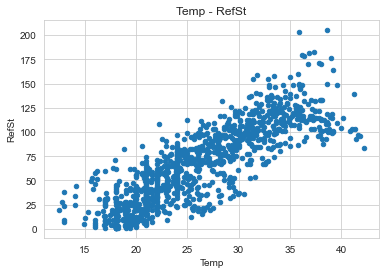
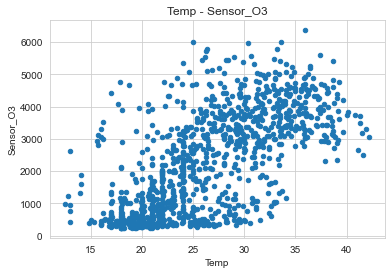


What we can conclude from the correlation matrix above? We can observe that the best features with the highest correlation are Sensor\_O3, Temp, ReHum and Sensor\_NO2, so these features are the one which resume better the value of the reference station. This information will be useful after in this report when we’ll do the feature selection.

But to avoid fast conclusions, let’s go to analyze the plots.

## Plots

### The temperature plots

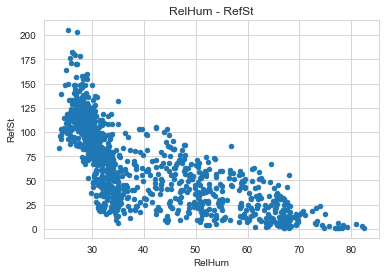
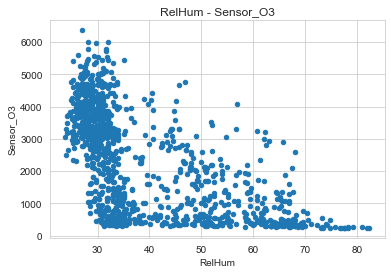
 

The two images on the tops are the plots about the temperature and the value of the O3/Reference station.

As we can see, the shapes in the plots is a little similar, but in the first one is more defined. Always in the first plot, is clearly visible that there is a linear evolution of the data, a thing also visible in the O3 plot but with some noise.

So, we can conclude that the temperature is one of the features that allow to compute the RefSt value in the linear regression.

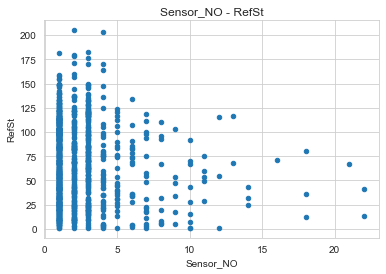
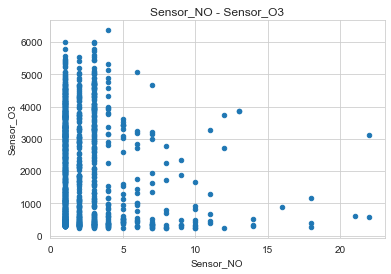
### Humidity plots

Also in this case, we can see that the plots above have more or less the same shape, with more sparsity in the second one. We remember that the humidity is the best negative correlation for RefSt in the negative matrix, this is visible in the plot because the data doesn’t follow any evolution and are concentrated in a region of the plot (in this case when the umidity is around 30). On the other hand, these is another data concentration in the second plot: it is in the low part, where the value of O3 is less than 1000. So, the humidity can be resumed in a hyperbole in the last plot.

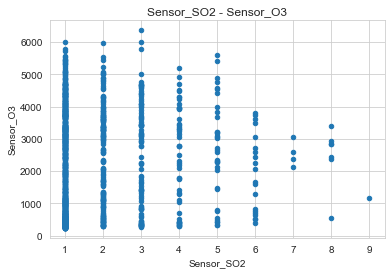
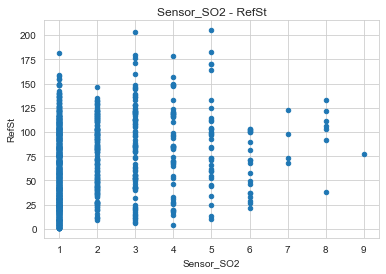
To make some conclusions, Surely the humidity doesn’t impact significantly in the linear regression, but it could inpact in the anothers. We’ll see it later.

### Nitrogen monoxide plots

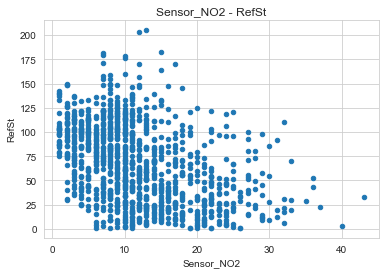
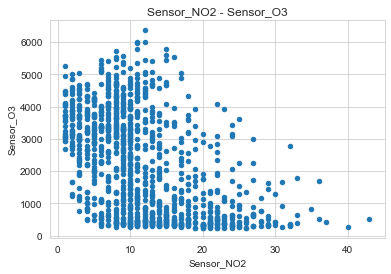
In these plots the data are concentrated in the first values. So, we can make more or less the same conclusions: the NO will impact less in a linear regression and probably it impact more or less at the same way with other algorithms because the correlation is very near to zero respect to the others.

### Sulfur dioxide plots



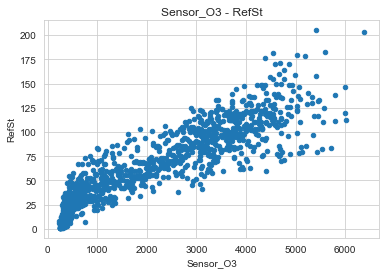
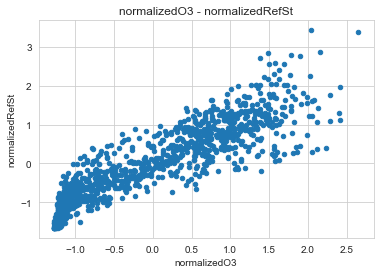
In this situation, we can make the same make the same conclusions of the previous case: because of this feature as a low correlation with RefSt and the plots has not a defined shape, The SO2 is not a good estimator of RefSt.

### Nitrogen Dioxide plots

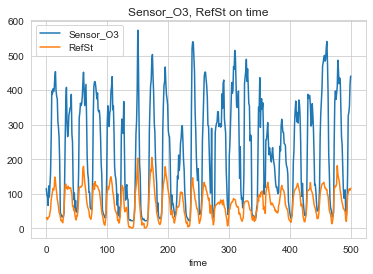
 

This is one of the features we chose from the correlation matrix, we can confirm it because the two plots above present similar shapes. The evolution is not linear, so this feature could be relevant in other algorithms.

### O3 plots

Plotting the O3 values respect to the reference station allows us to see the linear evolution of the last feature. This is a characteristic of an high correlation and we can confirm it with a value of 0.9 on the correlation matrix.



Another confirmation is the plot above, in fact it is clearly visible how the O3 and the reference station follow the same evolution.

NOTE: In order to see this similarity more clearly, the O3 data has been scaled of a factor 50, but it does not affect the evolution!

In order to make conclusions, Sensor\_O3 is the best feature to summarize the value of the reference station!

So, for summarizing everything:

* Sensor\_O3 and Temp are the best features!
* Sensor\_NO2 and RelHum are good features!
* Sensor\_NO and Sensor SO2 are bad features.

# Models

## Linear regression with subset selection

### Plots

## Lasso Linear Regression

### Tuning

### Plots

## Ridge Linear Regression

### Tuning

### Plots

## KNN Regression

### Tuning

### Plots

## Kernel Regression

### Tuning

### Plots

## Random forest regression

### Tuning

### Plots

## Support vector regression

### Tuning

### Plots

Fourth homework of TOML

# Introduction

# Case of overfitting

## Plot

# Case of underfitting

## Plot

# Right values for the model

## Plot

# Problems

## Plot