Themes on machine learning (TOML)

Third project

Contenido

[Introduction 2](#_Toc106305561)

[Data observation and first graphs 2](#_Toc106305562)

[Modeling and calibration 4](#_Toc106305563)

[Forward Subset Selection (MLR- FSS) 4](#_Toc106305564)

[Ridge Regression (MLR-RR) 4](#_Toc106305565)

[Lasso Regression (MLR-LR) 4](#_Toc106305566)

[K-nearest neighbor (KNN) 4](#_Toc106305567)

[Kernel Ridge Regression with RBF kernel 4](#_Toc106305568)

[Random Forest (RF) 4](#_Toc106305569)

[Support Vector Regression (SVR) 4](#_Toc106305570)

[Comparison and conclusions 4](#_Toc106305571)

# Introduction

The main goal of this project is to calibrate an air pollution sensor in an air pollution monitoring sensor network. To do this the data obtained by this network is modeled using different methods. The results will be compared using graphs and metrics and one of them will be selected as the best method in the conclusions part of the report.

The distribution of the report is as follows: first the sensor data files and data is explained and analyzed using graphs, then several modeling techniques are applied to this data and, finally, the conclusion of which is the best modeling technique is made using graphs and metrics to compare in the last section.

# Data observation and first graphs

At the beginning various data files must be merged into a single data file to enable the modeling in the next steps. The final merged file has the following columns containing data from different sensors:

* **date:** This column has the timestamp of the data.
* **RefSt:** This column has the O3 value of the reference station. The value is in µg/m³.
* **Sensor\_O3:** This column has the data of the O3 sensor in KOhms.
* **Temp:** This column has the temperature data in ºC.
* **RelHum:** This column has the relative humidity data in %.
* **Sensor\_NO:** NO sensor data.
* **Sensor\_NO2:** NO2 sensor data.
* **Sensor\_SO2:** SO2 sensor data.

First, let us compare the plots of the reference station data and the O3 sensor data. The magnitude should be different, but the shape should be the same since the data is the same and is taken in the same periods/intervals of time.

Gráfico, Gráfico de líneas

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamente

The shape of both graphs is the same. It has slight differences but overall, it has the same shape. The next step is to plot the relationship between both parameters. As mentioned before, both parameters are the same kind of data but taken by different sensors at the same time and with different units, thus, the relationship should give us a scatter plot with a 45 degrees shape approximately.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

The graph confirms that the data of both sensors is related and has the same kind of variations.

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

For the plotting of the former graphs, the data is not normalized since the analysis done in this section is more focused in the shape rather than in the magnitude or the value of the variables. For the modeling part, all the data is normalized due that is a mandatory previous step to get correct models.

# Modeling and calibration

## Forward Subset Selection (MLR- FSS)

## Ridge Regression (MLR-RR)

## Lasso Regression (MLR-LR)

## K-nearest neighbor (KNN)

## Kernel Ridge Regression with RBF kernel

## Random Forest (RF)

## Support Vector Regression (SVR)

# Comparison and conclusions