POLITECNICO
MILANO 1863

Politecnico di Milano

Scuola di Ingegneria Civile,Ambientale e Territoriale Master in Geoinformatics

# My Thesis

Bresciani Matteo

Referent professor: Brovelli Maria Antonia

**Abstract**

qq

# Contents

# Chapter 1

# Introduction

Science demonstrated that deterioration of ambient air quality, due to the growing concentration of pollutant in the atmosphere, has caused a significant increment of deaths in the world.

Pollutants such as particulate matter, ozone, carbon monoxide and ammonia cause respiratory diseases, and are important sources of mortality. Almost the entire global population (99%) breathes air that exceeds WHO air quality limits, and threatens their health.
In Europe air is getting cleaner, but persistent pollution, especially in cities, is damaging people's health. One of the last reports which is based on the European Environment Agency's (EEA), shows that exposure to air pollution caused around 500,000 early deaths in the European Union (EU) in 2018 [1].

One of the most harmful pollutants is the **particulate matter (PM25 or PM10)** which can get deep into your lungs or even get into your bloodstream.
Most of the particles come from chemical reactions such as sulphur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industries and automobiles.

However, a significant sources of PM are the chemical reactions generated by intensive farming [2]. In particular this is a relavant issue in the Po Valley, where intensive agricultural activity is very employed.

In this context, human civilization is trying to limit pollution and improve environment with use of technology. Technology is helping to clean up air pollution, with data analytics-based solutions helping to make our cities healthier places to live.
Monitoring, analysing and predicting the air quality in urban areas is one of the effective solutions for coping with the climate change problem.

The advent of modern Artificial Intelligence (AI) techniques such as Machine Learning (ML) can be considered as new possibilities for researchers to find solutions to various problems affecting air quality and climate change. In this context, the **D-DUST project** (Data-driven moDelling of particUlate with Satellite Technology aid),

funded by Fondazione Cariplo's 'Data Science for Science and Society' call for proposals, counts on Politecnico di Milano, Department of Civil and Environmental Engineering (DICA) as lead partner.
D-DUST aims to provide knowledge about the impact of agricultural and livestock activities on pollutants in the Po Valley (Northern Italy).

For reaching the goal, data from ground sensor are combined with contribution provided by satellite platforms and, with the use of data science techniques like machine-learning and geostatistical models, provide meaningful information related to the contribution of intensive farming on pollution.

The last target of the project is to provide a data-driven best-practices to policymakers, farming operators and citizens in order to minimize the production processes' effects on air quality. In this thesis we propose an

ensemble approach for analysing data and provide useful information regarding intense agricultural activity through selection of the most remarkable covariates that impact on PM25 and NH3 pollutants. The final step is to build a model skilled to estimate pollutant estimation locally, better than global scale model.

# Chapter 2

# Background and state of art

In this chapter I'm going to contextualize the state of the art of my research work. Besides, I'll give explanation about the target to reach and the solution applied.

The goal of my research is to implement a Machine Learning model capable of predicting pollutant locally with better precision than global scale. Data must be preprocessed before training, in order to reduce overfitting and improve accuracy of the final model. Another aspect to take in consideration is that a ML model trained with so many features it would be a black box, in which a lack of interpretability couldn't be able to explain the decisions taken by the AI. So it's needed to care about interpretability in order to discard eventually confoundant variables which can suggest there is a correlation when in fact there is not, even if the model's accuracy is extremely high. For instance, a new paper by Alex DeGrave et al.[3] shows that Deep Learning model trained with improper data was taking shortcuts in COVID-19 detection on radiographs because of position of certain markers rather than on the actual radiograph. Therefore, the key to increase interpretability of a given model is to wonder if given factor should drive the final decision.

In this context, in which the black-box nature of ML algorithms raises ethical and judicial concerns inducing lack of trust, Explainable Artificial Intelligence aims to create a model fully interpretable. Explainable Artificial Intelligence (or Explainable Machine Learning) helps to understand how ML algorithms make prediction, with the usage of Feature Selection methods for determining how well each feature can predict the target variable. Indeed, before developing a predictive model, feature selection is essential for reducing the number of input variable. It is desirable to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Nowadays, with the large amount of volume and variety in Big Data, FS is becoming increasingly an essential preprocessing step in machine learning algorithms [4].

In litherature we can find various feature selection methods which are classified in three main categories[6]:

### Filter Methods

Filter-based feature selection methods adopt statistical measures to evaluate the correlation/dependence between input variables. These select features from the without machine learning algorithm. In terms of computation, they are very fast and are very suitable in order to remove duplicated, correlated, redundant variables. On the contrary, these methods do not remove multicollinearity

### Wrapper Methods

Wrapper methods, as the name suggests, wrap a machine learning model, with different subsets of input features. In this way the subsets are evaluated following the best model performance. One disadvantage of this approach is the computational costs. Their exection for many subsets of variables can become unfeasible.

### Embedded Methods

Embedded methods instead are characterised by the benefits of both the wrapper and filter methods, by including interactions of features but also having a reasonable computational cost.

Due to the fact there isn't a best feature selection technique, I performed and combined different supervised methods. According to the above, my work comes on this scenario, having the aim to preprocessed geospatial data in order to highlight the most weighted input variable that affect the pollutants as target variable. In the next chapter of my report details I'll show the tools developed and the strategy choosen to reach the target. This is the content of the next chapters:

- chapter 3 (**Overview**): It shows the main steps I take in my work;
- chapter 4 (**Data Collection and Pre-Processing**): It outlines in detail how data are collected and preprocessed, with particular attention to the feature selection;
- chapter 5 (**Case of Study and Data Modelling**): It's focused on the results achieved in the case of study (both feature selection and models built);

# Chapter 3

# Overview

## 3.1 Pre-processing

My work is focused on the first phase of a data analysis procedure which is the pre-processing. Data pre-processing (or data preparation) is the process of transforming raw data into a suitable format for modelling. Indeed, raw data is in most cases incomplete and noisy.

Nowadays, dealing with big amount of information, the probability of incorrect data is higher without a proper data pre-processing. Only high-quality data can generate accurate models and predictions.

Hence, it's crucial to process data with the best possible quality before training them with artificial intelligence, and machine learning predictive models.

For doing this I implemented tools collected in Python Notebooks, each one available in the D-DUST repository: (`https://github.com/opengeolab/D-DUST/tree/thesis_MB`).
Its essential steps (shown in figure 3.1) are these.

### 3.1.1 Data Collection

Relevant data is gathered from their sources and merged in data structures (such as Dataframes). In our work, data come from fixed ground-sensor, satellite-based platform, models and map layers. In this phase are processed (mostly) numerical and categorical data.

### 3.1.2 Data Cleaning

It involves fixing problems or errors in messy or incomplete data. There are general data cleaning operation, such as identifying:

- duplicate rows of data and remove them;
- rows with NaN values and remove them;
- columns that have low variance and drop them;

### 3.1.3 Data Transformation

Data need to be scaled. As a matter of fact, each feature in our data has varying degrees of magnitude, range, and units. This is an issue for machine learning algorithms beacuse of highly sensitive to these features. So in input or output data we performed:

- Standardization: Scale a feature to a standard Gaussian distribution;

- Normalization: Scale a feature to the range between 0 and 1;

### 3.1.4 Feature Selection

Feature Selection is the core part of this study. It's the process of reducing the number of input variables when developing a predictive model. Data collected, even if have been cleaned and transformed, are anyway characterized by big amount of variables which are redundant. Discarding irrelevant data is essential before applying Machine Learning model in order to:

- **Reduce Overfitting**: less opportunity to make decisions based on noise;
- **Improve Accuracy**: less misleading data means that modelling accuracy improves. Predictions can be greatly distorted by redundant attributes;
- **Reduce Training Time**: With less data an algorithm will train faster;

In this step, which will be explained in detail in the next chapters, the reduced input variables are the ones that are meaningless with respect to a target variable as output.
In this study target variables chosen represent the pollution phenomena such as the PM25 and Ammonia emissions. We choose these targets because are the most relevant sources of pollution produced by intensive agricultural.
One of the aim of this step is to detect main pollutant factors which contribute further on the training of PM25 or NH3 emissions. Due to the fact that there isn't a best feature selection technique, many different methods are performed, each one that give different correlation results.

After this step, for every method, a score evaluation is assigned to each variable representing its contribution on the output.

Finally a voting algorithm is performed in order to average the scores obtained in each feature selection method. The highest values are selected for model as input.

## 3.2 Model prediction

Prediction is a type of analysis that uses techniques and tools to build predictive models and forecast outcomes. In my work predictive analysis is performed for making prediction on pollutants with data processed in the first phase as input.
Model predictions are deployed through regression analysis, used for estimating the relationships between a dependent variable and one or more independent variables.

In particular I use supervised techniques based on Machine Learning where the model built is fit with the training dataset and evaluated its performance with the testset. For doing prediction, I employ 2 supervised AI models:

- **Neural Network regression with Keras**: It's one of the deep learning algorithms that simulate the workings of neurons in the human brain. In a neural network neurons are linked between them forming layers;
- **Machine Learning with Random Forest regressor**: It operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees;

After this step an evaluation of the performance of predictions is performed in terms of error and accuracy with a procedure called k-fold cross validation.
Finally, a comparison with CAMS data is performed with the aim to demonstrate that the models produced are better estimated in this local scale.

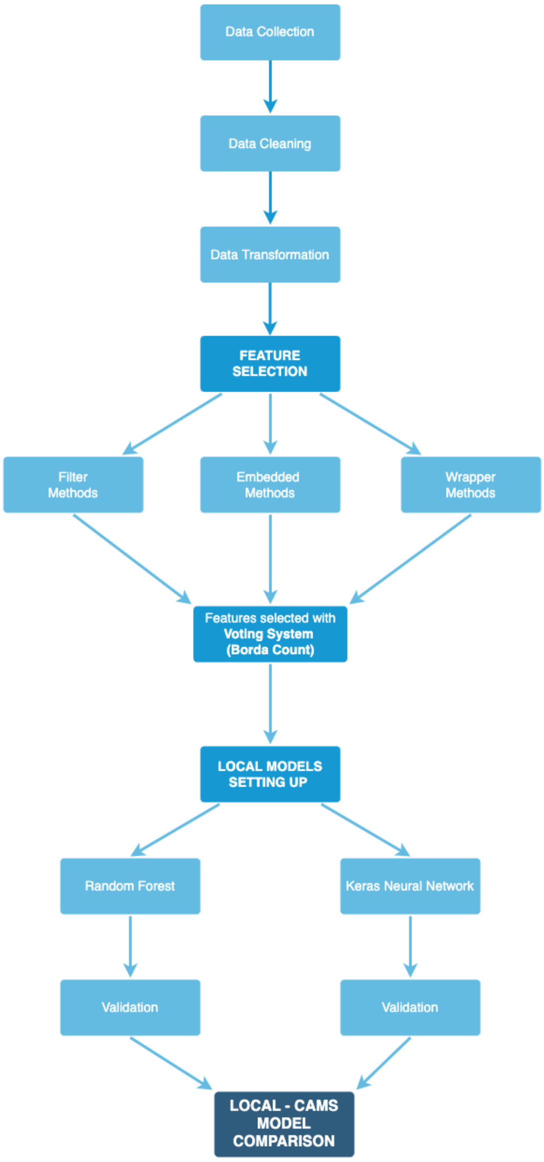


Figure 3.1: Overview of the steps made.

In the next chapters each step will be described in depth about procedures adapted and results obtained.

# Chapter 4

# Data Collection and Pre-Processing

In this chapter I explain each step taken during the preprocessing phase, by illustrating in details each step taken and tool used.

## 4.1 Data Collection

Data collection is the process of gathering information in variables of interest for answering relevant questions. Variables selected are the physical and chemical factors that are most associated with the formation of primary and secondary pollutant. Therefore, the variables are categorized in 4 different labels:

- Weather: These elements, such as wind speed and direction, precipitation and air temperature, changes in the epochs and can influence air pollution;
- Pollutant: These variables represent primary and secondary pollutant related to the greenhouse effect;
- Soil and Vegetation: Since soil and vegetation degradation are global concerns and can inluence the air propagation in the environment, data related to local morphology are collected;
- GIS (static layers): This time-invariant layers are considered to be changeless in the time range considered. Differently from the other types which need a constant monitoring, these variables are update yearly with a lower frequency than the others;

Data chosen are open source and regularly available. In this phase data have been collected (not by me but by other collegues of D-DUST project at this link) in grids from different sources and provider in Geopackages.

### 4.1.1 Source types

In order to better distinguish the data sources characteristics, variables selected are labeled with 4 different types:

- Ground Sensor: Each ground monitoring stations belongs mainly to ARPA and ESA provides meterological and air quality data;
- Model: data are estimated through a model built using satellite and meteorological and air quality data as input, such as european data provided by CAMS (Copernicus Atmosphere Monitoring Service);
- Map layer: this data are time-invariant and are related to Lombardy morphology such as density of roads, population or land use;
- Satellite Sensor: They provide data from air quality observation mainly. Satellites provider are Sentinel-5P Tropomi and Terra & Aqua MODIS;

### 4.1.2 Spatial resolution

Vector grids that are used in the D-DUST project are three and they are generated by the spatial resolution of the source provider.

- Grids with 0.1° resolution with Copernicus CAMS (European);
- Grids with 0.066° resolution based on S5P (this resolution is not included in the case of study);
- Grids with 0.01° Grid defined with maximum one ARPA station for each cell;

Data are scaled and fit in each spatial resolution grid in order to better analyze the final output model by considering each of them.

In the next lines each variable is provided in tables, by showing its type, name and description:

**Meteo**

| **Physical variable** | **Source type** | **Variable name** | **Description** | **Unit** | **Source** |
|---|---|---|---|---|---|
| Temperature | Model | temp_2m | Mean air temperature at 2 m above the land surface | K | ERA5-Land hourly data. |
| | Ground Sensor | temp_lcs | Mean air temperature ground measurement - Low Cost Sensor ESA monitoring stations. | °C | ESA Air Quality Platform. |
| | Ground Sensor | temp_st | Mean temperature - ARPA monitoring stations. | °C | ARPA Lombardia. |
| Wind | Model | e_wind | Mean eastward wind component 10 m above the land surface | m/s | ERA5-Land hourly data. |
| | Ground Sensor | wind_dir_st | Wind direction from ground sensor divided in 8 sectors. These are classified into 8 categories as specified in "Notes" column. | cat | ARPA Lombardia. |
| | Ground Sensor | n_wind | Mean northward wind component 10 m above the land surface. | m/s | ERA5-Land hourly data. |
| | Ground Sensor | wind_speed_st | Mean wind speed on ground - ARPA monitoring stations. | m/s | ARPA Lombardia. |
| Precipitation | Model | prec | Mean accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface. It is the sum of large-scale precipitation. | m | ERA5-Land hourly data. |
| | Ground Sensor | prec_st | Mean precipitation in each cell in the time range - ARPA monitor stations. | mm | ARPA Lombardia. |
| Air Humidity | Ground Sensor | air_hum_st | Mean air moisture measurement in the time range - ARPA monitoring stations | % | ARPA Lombardia. |

| | | | | | |
|---|---|---|---|---|---|
| | Ground Sensor | air_hum_lcs | Mean air moisture ground measurement - Low Cost Sensor ESA monitoring stations. | % | ESA Air Quality Platform. |
| Air Pressure | Model | press | Mean weight of all the air in a column vertically above the area of the Earth's surface represented at a fixed point. | Pa | ERA5-Land hourly data. |
| Solar Radiation | Ground Sensor | press | Global radiation measurement - ARPA monitoring station. | $W/m^2$ | ARPA Lombardia. |

**Pollutants**

| **Physical variable** | **Source type** | **Variable name** | **Description** | **Unit** | **Source** |
|---|---|---|---|---|---|
| | | | *Pollutants* | | |
| Dust | Model | dust | Mean dust concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | $ug/m^3$ | CAMS Model. |
| AOD | Satellite Sensor | aod_055 | Mean Aerosol Optical Depth at 550nm. | - | MODIS Terra+Aqua. |
| | Satellite Sensor | aod_047 | Mean Aerosol Optical Depth at 470nm. | - | MODIS Terra+Aqua. |
| | Satellite Sensor | uvai | Mean UV Aerosol Index. A positive index highlights the presence of UV absorbing aerosol (such as smoke/dust). | - | Sentinel-5P |
| PM10 | Model | pm10_cams | Mean PM10 concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | $ug/m^3$ | CAMS Model. |
| | Ground Sensor | pm10_lcs | Mean PM10 concentration ground measurement - Low Cost Sensor ESA monitoring stations. | ? | ESA Air Quality Platform. |
| | Ground Sensor | pm10_st | Mean PM10 concentration ground measurement - ARPA monitoring stations. | $ug/m^3$ | ARPA Lombardia |
| PM2.5 | Model | pm25_cams | Mean PM2.5 concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | $ug/m^3$ | CAMS Model. |

| | | | | | |
|---|---|---|---|---|---|
| | Ground Sensor | pm25_lcs | Mean PM2.5 concentration ground measurement - Low Cost Sensor ESA monitoring stations. | ug/m$^3$ | ESA Air Quality Platform. |
| | Ground Sensor | pm25_st | Mean PM2.5 concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia |
| $SO_2$ | Model | so2_cams | Mean $SO_2$ concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| | Satellite Sensor | so2_s5p | Mean SO2 vertical column density at ground level. | mol/m$^2$ | Sentinel-5P. |
| | Ground Sensor | so2_st | Mean $SO_2$ concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia. |
| $NO_2$ | Model | no2_cams | Mean $NO_2$ concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| | Satellite Sensor | no2_s5p | Mean NO2 vertical column density at ground level. | mol/m$^2$ | Sentinel-5P. |
| | Ground Sensor | no2_st | Mean $NO_2$ concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia. |
| | Ground Sensor | no2_lcs | Mean $NO_2$ concentration ground measurement - Low Cost Sensor ESA monitoring stations. | ug/m$^3$ | ESA Air Quality Platform. |
| NO | Model | no2_cams | Mean NO concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| $NO_x$ | Ground Sensor | nox_st | Mean $NO_x$ (field: "Ossidi di Azoto") concentration ground measurement - ARPA monitoring stations | ug/m$^3$ | ARPA Lombardia. |
| $CO_2$ | Ground Sensor | c02_lcs | Mean CO2 concentration ground measurement - Low Cost Sensor ESA monitoring stations | ? | ESA Air Quality Platform. |
| CO | Model | co_cams | Mean CO concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| | Satellite Sensor | co_s5p | Mean CO vertically integrated column density. | mol/m$^2$ | Sentinel-5P. |

| | | | | | |
|---|---|---|---|---|---|
| | Ground Sensor | co_st | Mean CO concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia. |
| | Ground Sensor | co_lcs | Mean CO concentration ground measurement - Low Cost Sensor ESA monitoring stations. | ug/m$^3$ | ESA Air Quality Platform. |
| $O_3$ | Model | o3_cams | Mean $O_3$ concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| | Satellite Sensor | 03_s5p | Mean $O_3$ total atmospheric column | mol/m$^2$ | Sentinel-5P. |
| | Ground Sensor | 03_st | Mean $O_3$ concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia. |
| $CH_2O$ | Satellite Sensor | ch20_s5p | Mean Formaldehyde tropospheric column number density | mol/m$^2$ | Sentinel-5P. |
| NMOVOCs | Model | nmvocs_cams | Mean Non-Methane VOCs concentrations at 0m level provided by CAMS. | ug/m$^3$ | CAMS Model. |
| $NH_3$ | Model | nh3_cams | Mean $NH_3$ concentration at 0m level provided by CAMS (Ensemble Median - Analysis). | ug/m$^3$ | CAMS Model. |
| | Satellite Sensor | nh3_lcs | Mean $NH_3$ concentration ground measurement - Low Cost Sensor ESA monitoring stations | ? | ESA Air Quality Platform. |
| | Ground Sensor | nh3_st | Mean $NH_3$ concentration ground measurement - ARPA monitoring stations. | ug/m$^3$ | ARPA Lombardia. |

**Soil and Vegetation**

| **Physical variable** | **Source type** | **Variable name** | **Description** | **Unit** | **Source** |
|---|---|---|---|---|---|
| | | | *Pollutants* | | |
| Vegetation | Satellite Sensor | siarlX | Fraction of area in each cell for each agricultural use provided by SIARL Catalog for Lombardy Region. | % | SIARL Lombardia 2019. |
| | Satellite Sensor | ndvi | Mean NDVI cell value over 16 days period | - | USGS Earth Data. |

| | | | | | |
|---|---|---|---|---|---|
| | Satellite Sensor | siarl | Majority class for agricultural use provided by SIARL Catalog for Lombardy Region. | cat | SIARL Lombardia 2019. |
| Soil | Model | soil_moist | Mean volume of water in soil layer 1 (0 - 7 cm) of the ECMWF Integrated Forecasting System. The surface is at 0 cm. The volumetric soil water is associated with the soil texture (or classification), soil depth, and the underlying groundwater level. | $m^3/m^3$ | ERA5 Land Hourly Data. |
| | Map Layer | soilX | Fraction of area for each cell containg the soil type obtained from OpenLandMap soil texture classification. | % | OpenLandMap Soil Texture Class (USDA System). |
| | Map Layer | soil_textX | Mean NDVI cell value over 16 days period | % | Basi informative dei suoli - Geoportale Lombardia. |
| | Map Layer | soil | Majority soil type for each pixel from OpenLandMap soil texture classification . | cat | OpenLandMap Soil Texture Class (USDA System) . |
| | Map Layer | soil_text | Majority soil type for each pixel from Carta pedologica 250K (Lombardy Region). | cat | Basi informative dei suoli - Geoportale Lombardia. |

**GIS (static layers)**

| **Physical variable** | **Source type** | **Variable name** | **Description** | **Unit** | **Source** |
|---|---|---|---|---|---|
| | | | *Pollutants* | | |
| Geometry | Map Layer<br>Satellite Sensor | area<br>ndvi | Area of Lombardy Region vector layer in each cell.<br>Mean NDVI cell value over 16 days period | $km^2$<br>- | SIARL Lombardia 2019.<br>- |
| Population | Map Layer | pop | Population for each cell. | n° of inhabitants | Gridded Population of the World (GPW). |
| Land use and cover | Map Layer | dsfX | Land use fraction for each cell containing the classification the classification provided by DUSAF Catalog (Lombardy Region). | % (fraction for each cell) | DUSAF Lombardia 2018. |

| | | | | | |
|---|---|---|---|---|---|
| Map Layer | dusaf | Cover | Land Use majority class for each cell provided by DUSAF Catalog (Lombardy Region). | cat | DUSAF Lombardia 2018. |
| Terrain | Map Layer | h_mean | DTM average elevation for each pixel. | m | Geoportale Lombardia 2019. |
| | Map Layer | aspect_major | Aspect derived from DTM. Majority pixel aspect. | Degree North | Geoportale Lombardia 2019. |
| | Map Layer | slope_mean | Average slope derived from DTM. | Degree North | Geoportale Lombardia 2019. |
| Road Infrastructures | Map Layer | int_prim | Density of intersection nodes between primary roads for each cell (including highways). | $\text{int}_s/\text{km}^2$ | Geoportale Lombardia 2019. |
| | Map Layer | int_prim_sec | Density of intersection nodes between primary and secondary roads for each cell. | $\text{int}_s/\text{km}^2$ | Geoportale Lombardia 2019. |
| | Map Layer | int_sec | Density of intersection nodes between secondary roads for each cell. | $\text{int}_s/\text{km}^2$ | Geoportale Lombardia 2019. |
| | Map Layer | prim_road | Density of primary importance roads for Lombardy Region inside for each. | $\text{km}/\text{km}^2$ | Geoportale Lombardia 2019. |
| | Map Layer | sec_road | Density of secondary importance roads for Lombardy Region foreach cell. | $\text{km}/\text{km}^2$ | Geoportale Lombardia 2019. |
| | Map Layer | highway | Density of highways for Lombardy Region inside for cell divided. | $\text{km}/\text{km}^2$ | Geoportale Lombardia 2019. |
| Farms | Map Layer | farms | Fration of area covered by farms inside the cell. Obtained from DUSAF dataset. | % (fraction for each cell) | DUSAF Lombardia 2018. |
| Air quality zones | Map Layer | aq_zone | Majority class of a given air quality zone in each cell. | cat | Geoportale Lombardia. |
| Climate zones | Map Layer | clim_zone | Majority class of a given air quality zone in each cell. | cat | - |

**Categorical Variable**

Categorical data are identified with names or labels given to them as value. Even if are rapresented by numbers, they don't have the same mathematical meaning as a numerical value. This type of data is discarded during the preprocessing phase, since feature selection is done exclusively on numerical input and output values. In the following table is explained the semantic of the values assumed.

| Variable name | Note |
|---|---|
| *Meteo* | |
| wind_dir_st | 1 = North: 0° - 22.5° / 337.5° - 360°,<br>2 = North-East: 22.5° - 67.5°,<br>3 = East: 67.5° - 112.5°,<br>4 = South-East: 112.5° - 157.5°,<br>5 = South: 157.5° - 202.5°,<br>6 = South-West: 202.5° - 247.5°,<br>7 = West: 247.5° - 292.5°, 8 = North-West: 292.5° - 337.5° |
| *Soil and Vegetation* | |
| siarl | 2 = Cereal<br>9 = Mais<br>12 = Rice |
| soil | 2=Cereal<br>9=Mais<br>12=Rice |
| soil_text | 1 = Clay<br>2 = Silty Clay<br>3 = Sandy Clay<br>4 = Clay Loeam<br>5 = Silty Clay Loam<br>6 = Sandy Clay Loam<br>7 = Loam<br>8 = Silt Loam<br>9 = Sandy Loam<br>10 = Silt<br>11 = Loamy Sand<br>12 = Sand. |
| *GIS (Static layers)* | |
| dusaf | 2 = Agricultural areas<br>3 = Wooded territories and semi-natural environments<br>4 = Wetlands<br>5 = Water bodies<br>11 = Urbanised areas<br>12 = Production facilities, large plants and communication networks<br>13 = Mining areas, landfills, construction sites, waste and abandoned land<br>14 = Non-agricultural green areas |
| aq_zone | 1 = Highly urbanized plains<br>2 = Plains<br>3 = Prealpi, Appennino and mountains<br>4 = Valley floor Agg.<br>5 = Urban agglomarated area (Milano, Bergamo, Brescia). |
| clim_zone | 1= Alpi<br>2 = Prealpi Occidentali<br>3 = Prealpi Orientali<br>4 = Pianura Occidentale<br>5 = Pianura Centrale<br>6 = Pianura Orientale. |

## 4.2 Data Cleaning

Data has to be prepared in accordance with the supervised feature selection. Data cleaning aims to fix problems or errors in messy data. There are many reasons data may have incorrect values, such as being corrupted, duplicated or invalid. This could be done by removing a rows or columns. Alternately, it might involve replacing observations with new values. Firstly data covariates are splitted between independent (X) and target variable (y). X represents all of variables collected in the previous part, exceting for the pollutant to be analyzed and modelled (such as PM25 or Ammonia) which consists in the y variable.

In this section are underlined the issues mitigated by the Data Cleaning.

### 4.2.1 Nan Values

In my work I consider as pollutant variables coming from ground sensor measurament. Air quality monitoring is usually carried out through ground sensors networks, which represents the primary air quality data source by governance. However, no country in the world has yet established a monitoring network with a full satisfying coverage[5]. Even in the United States (US), which is characterized by a relatively developed PM2.5 ground monitoring network with 2500 stations has many areas unmonitored[5]. It implies that in our dataset collected there are many samples with NaN values due to the features provided by ground sensor.

It's feasible since:

- A sensor could have no measurement for a given time epoch;
- The set of sensor, because of its limited supply, cannot cover each cell of a grid;

In our case variable provided by ARPA and ESA ground sensor (with the label that ends with '_st' and '_lcs' respectevively) has many NaN cells.

In order to mitigate this, I present this solution in sequence:

- Drop of samples having target variable with NaN value;
- Drop of columns having at least a NaN value;

Due to the fact that it results a dataset with a very limited number of sample [8], I perform additionally a k-nearest neighbor classifier[7] for adding a buffer of values close to the location of the ground stations measurement. In this way the size of the final sample, as the performance of the feature selection would increase.

In the grids processed there's the problem that a given value provided by measurement tools (such as ground and satellite sensor) could be NaN.

## 4.3 Data Transormation

## 4.4 Feature Selection

# Chapter 5

# Case of Study and Data Modelling

# Chapter 6

# Conclusion

# Chapter 7

# Bibliography


[1] European Environment Agency. Air quality in europe 2018. *Report No. 12/2018*, 2018.

[2] MR Burkart. Diffuse pollution from intensive agriculture: sustainability, challenges, and opportunities. *Water science and technology*, 55(3):17–23, 2007.

[3] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[4] Saidjon Kamolov. Feature selection: state-of-the-art survey. *Annals of Mathematics and Computer Science*, 4:48–54, 2021.

[5] Ying Liu, Guofeng Cao, Naizhuo Zhao, Kevin Mulligan, and Xinyue Ye. Improve ground-level pm2. 5 concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution*, 235:272–282, 2018.

[6] Urszula Stańczyk. Feature evaluation by filter, wrapper, and embedded approaches. In *Feature Selection for Data and Pattern Recognition*, pages 29–44. Springer, 2015.

[7] Kashvi Taunk, Sanjukta De, Srishti Verma, and Aleena Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260. IEEE, 2019.

[8] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, 4(1):1–8, 2018.