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

July 15, 2022

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.

1.1 D-Dust

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

Overview

2.1 Pre-processing

Mainly my work (2.1) 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. Its essential steps are these.

2.1.1 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* and *satellite-based platform* and are entirely numerical.

2.1.2 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;

2.1.3 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;

2.1.4 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.

2.2 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 validation procedure.
This is done following a k-fold cross validation procedure. Finally, a comparison with CAMS data is performed with the aim to demonstrate that the models produced are better estimated in this local scale. In the next chapters each step will be described in depth about procedures adapted and results obtained.

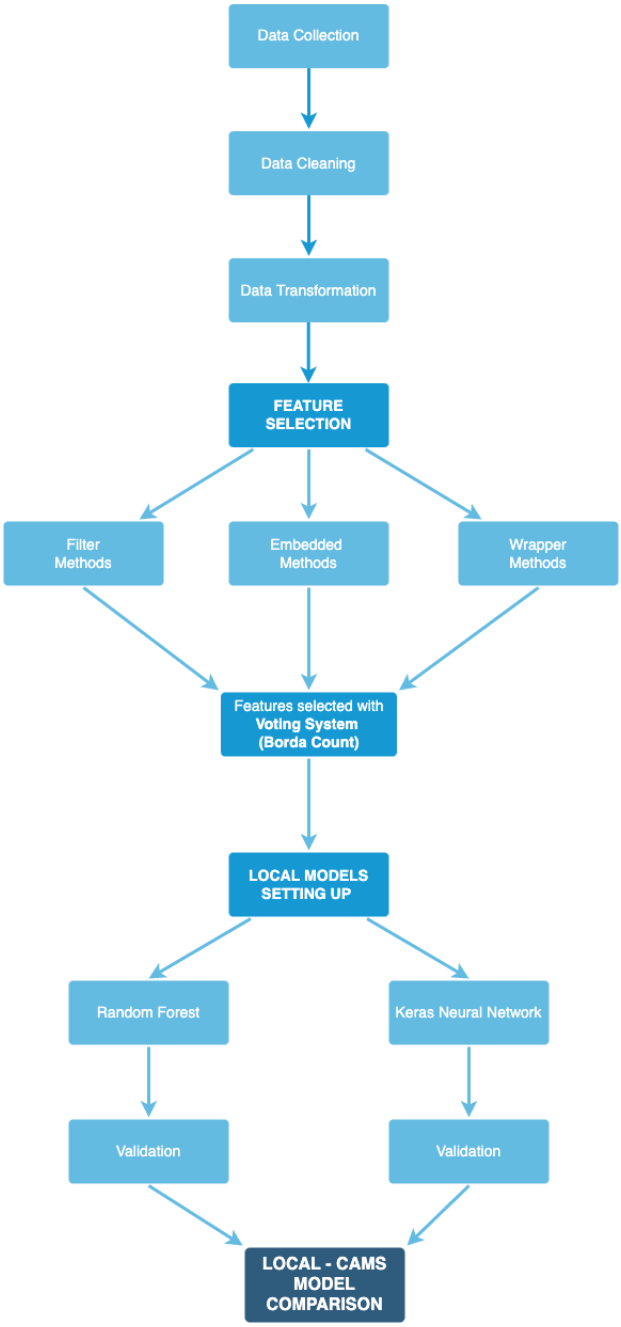

Figure 2.1: Overview of steps made.

Chapter 3

Background and state of art

Chapter 4

Data Collection and Pre-Processing

Chapter 5

Case of Study (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.