Study and prediction of air quality in smart cities through machine learning techniques

Ditsuhi Iskandaryan
Universitat Jaume I, Spain
iskandar@uji.es

Abstract

To support sustainable development, to regulate and reduce air pollution, it requires to implement appropriate measurements and approaches. One of these approaches will be designed and evolved in the scope of this research. It focuses on applying machine learning techniques for the purpose of performing air quality prediction with higher accuracy for different features: prediction target, data rate, period, algorithm, time granularity, evaluation metric, by using different datasets related to air quality.

1 PhD research topic, aim and objective

The thesis, which will be developed within the framework of the PhD, is titled as 'Study and prediction of air quality in smart cities through machine learning techniques'. Considering the impact of air quality on the environment and health, it becomes more prominent to monitor and control air quality. The smart city concept by integrating Information and Communication Technology (ICT) with citizens and existing resources fosters to achieve the above-mentioned goal. Specifically, it is noteworthy the role of machine learning technologies, which enables to predict and make decisions based on the collected data. Therefore, the principal objective of the thesis is to predict air quality effectively and efficiently in smart cities by implementing machine learning approaches; to create a framework based on the same conditions in order to validate and compare the advancement of different algorithms with different dataset types.

In: A. Kmoch, E. Uuemaa, D. Nüst (eds.): Proceedings of the 5th AGILE PhD School 2019, Tartu, Estonia, November 2019, published at https://doi.org/10.5281/zenodo.3835767.

2 Visualization

After finding a proper method or methods to predict air quality, an additional step can be to visualize the outputs with the intention of making the results available for all levels of the society. One of the objectives will be to display spatial-temporal changes of air quality index in the city for the next few days. The product will allow the users to select time granularity from the available options. Different graphs and charts will be drawn to show connections of different features.

3 Reproducibility

Taking into account the contribution of reproducibility in science nowadays, several techniques will be used to make the research reproducible.

The scripts and detailed description of methodology (software, environmental variables, platform dependencies) implemented in the work will be available at GitHub repository (https://github.com/Ditsuhi).

As a programming language, mostly Python with Google Colab (https://colab.research.google.com) will be used. Colab combines texts with documentations, figures and tables using markdown. It allows to create, upload, store and share notebooks, link to GitHub profiles and upload notebooks from GitHub, contributing transparency and reusability.

To provide reproducibility of complex workflows and data created and developed in the research, Docker containerisation tool (https://www.docker.com/) and Zenodo repository (https://www.zenodo.org/) will be used.

Data

The case study of the research is the city of Madrid, and the data used for analysis are classified as follows:

Air quality data: are available on-line by this link: https://datos.madrid.es/portal/site/egob/. These data are provided by the General Directorate of Sustainability and Environmental Control of Spain. They have started to capture data since 2001. There is additional information describing and interpreting the data. Apart from the archive, it includes hourly and daily data, it is also available the data captured in real time.

Meteorological data: includes temperature, humidity, wind direction, wind speed and so on, which will be obtained from AEMET OpenData with the following address: `https://opendata.aemet.es` (AEMET is a public body currently attached, through the Secretary of State for the Environment, to the Ministry of Agriculture, Food and Environment);

Spatial data: are the locations of the stations;

Temporal data: contains information indicating day of the month, day of the week, and the hour of the day in order to consider the overloaded duration from not overloaded;

and **Traffic data**.

4 Science Communication

The target audience of this research can be categorised into the following groups:

1) Scientists focused on air pollution, monitoring changes, estimating, forecasting, but mainly focusing on the air pollution aspect.

2) Scientists focused on machine learning technologies, to see how machine learning can be applied in a certain domain, how it handles specific types of data and find out the advantages and disadvantages of the applied techniques by comparing to each other.

3) Authorities of a city - the results can guide authorities of a city to apply protective measures in order to reduce air pollution and improve citizens health.

4) Each citizen in the city - taking into consideration adverse health effects caused by air pollution (increased risk of asthma, heart attack, cardiopulmonary and lung cancer, etc.), the information about air quality prediction and advance alerts can be beneficial for citizens to organise their daily activities by taking safety measures and preventing contaminated areas.