Statistical Learning Project

1st Milestone

G08

Prediction and analysis of air quality

Riccardo Ceccaroni, Giusy Beatrice Colarusso, Gabriele Giannotta, Francesco Lauro

Abstract

Air pollution is one of the big issues that affect metropolitan areas all around the world. Traffic and factories play a significant role. We need to implement models that keep track of data about the concentration of air pollutants (SO_2, NO_2, etc.). Indeed, the diffusion of these toxic gases in the air affects people's quality of life. The increase in data availability and new discoveries in terms of computational technologies are making it possible to predict and analyse the air quality, providing extremely useful information to control air pollution.

Main research aim & framework

The main goal of the project is to find a model that manages to predict air quality. Our interest in this topic stems from the fact that we think it's important to know more about air quality situation. Indeed, we strongly believe that predictions could help reducing pollution. Living in a "unique" historical period, namely the Covid-19 pandemic, a secondary goal of our research is understanding if this virus, besides changing our habits, has had an effect on air quality.

Data source(s)

Data will be collected on a daily basis (about 4 records per day) from the Air Quality Programmatic APIs website. They will be collected into a dataframe through our own Python script. This script connects directly to the website using APIs, and requests the air quality data measured in any data-gathering entity in the city of Rome. Each record will be used to fill in a dataframe which will be saved into a .csv format file. Finally, all these dataframes will be merged together.

Data collection

From each data-gathering station we will collect the following data: date, time, location, AQI index, carbon monoxide CO, nitrogen dioxide NO_2, ozone O_3, sulphur dioxide SO_2, particulate matter PM_{10}, particulate matter $PM_{2.5}$, humidity h, pressure p, temperature t, and wind speed w.

The main concern in recording data is that data-gathering entities are uploaded at a different time during the day, and some of them are updated only once a day. For this reason, we could be having some stations with more recordings than others.

Model & Methods

The two models we want to test are the following:

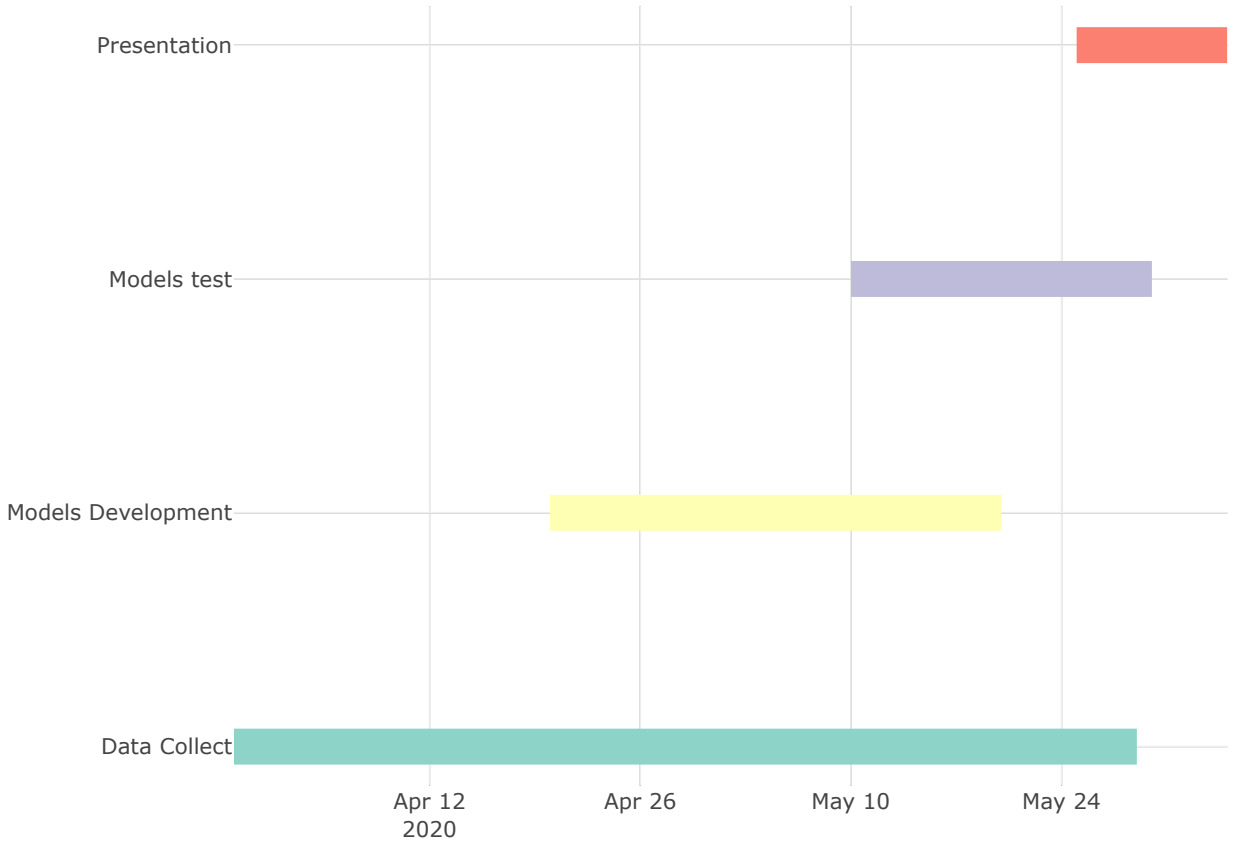
1. **Multiple Additive Regression Trees**: a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees (Wikipedia 2020b), (Karimian et al. 2019).
2. **ARIMA Model**: a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (Wikipedia 2020a), (Bhalgat, Bhoite, and Pitare 2019).

Software/Hardware Toolkit

Up to this point, the only programming language we have been using is *Python*.

It will be the main language we are going to use for the entire data analysis, together with *R*. Moreover, we will probably employ gbm and stats.

Project Timeline



References

Bhalgat, Pooja, Sachin Bhoite, and Sejal Pitare. 2019. "Air Quality Prediction Using Machine Learning Algorithms." *International Journal of Computer Applications Technology and Research* 8 (September). https://doi.org/10.7753/IJCATR0809.1006.

Delavar, Mahmoud, Amin Gholami, Gholam Shiran, Yousef Rashidi, Gholam Nakhaeizadeh, Kurt Fedra, and Smaeil Afshar. 2019. "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran." *ISPRS International Journal of Geo-Information* 8 (February): 99. https://doi.org/10.3390/ijgi8020099.

Karimian, Hamed, Qi Li, Chunlin Wu, Yanlin Qi, Yuqin Mo, Gong Chen, Sonali Sachdeva, and Xianfeng Zhang. 2019. "Evaluation of Different Machine Learning Approaches in Forecasting Pm2.5 Mass Concentrations." *Aerosol and Air Quality Research* 19 (January). https://doi.org/10.4209/aaqr.2018.12.0450.

Kleine Deters, Jan, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk. 2017. "Modeling Pm 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters." *Journal of Electrical and Computer Engineering* 2017 (June): 1–14. https://doi.org/10.1155/2017/5106045.

Qi, Zhongang, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li, Zhongfei, and Zhang. 2017. "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality." *IEEE Transactions on Knowledge and Data Engineering* PP (November). https://doi.org/10.1109/TKDE.2018.2823740.

Venegas, Laura, Nicolás Mazzeo, and Mariana Dezzutti. 2014. "A Simple Model for Calculating Air Pollution Within Street Canyons." *Atmospheric Environment* 87 (April): 77–86. https://doi.org/10.1016/j.atmosenv.2014.01.005.

Wang, Ping, Yong Liu, Zuodong Qin, and Guisheng Zhang. 2014. "A Novel Hybrid Forecasting Model for Pm10 and So2 Daily Concentrations." *The Science of the Total Environment* 505C (November): 1202–12. https://doi.org/10.1016/j.scitotenv.2014.10.078.

Wikipedia. 2020a. "Autoregressive integrated moving average — Wikipedia, the Free Encyclopedia." http://en.wikipedia.org/w/index.php?title=Autoregressive%20integrated%20moving%20average&oldid=944420432.

———. 2020b. "Gradient boosting — Wikipedia, the Free Encyclopedia." http://en.wikipedia.org/w/index.php?title=Gradient%20boosting&oldid=947218348.