Statistical Learning Project

1st Milestone

G08

Prediction and analysis of air quality

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

Abstract

L'inquinamento atmosferico è uno dei grandi problemi che affligge le aree metropolitane di tutto il mondo. Il traffico e le industrie svolgono un ruolo significativo. Abbiamo bisogno di implementare modelli che registrino informazioni sulle concentrazioni di inquinanti atmosferici (SO_2, NO_2, ecc.) poichè la deposizione di questi gas nocivi nell'aria sta influenzando la qualità della vita delle persone.

La crescita della disponibilità di dati e l'avanzamento delle tecnologie computazionali stanno rendendo possibile la previsione e l'analisi della qualità dell'aria, fornendo informazioni estremamente utili per controllare l'inquinamento atmosferico.

Main research aim & framework

Lo scopo principale del progetto è trovare il modello migliore che riesca a predire la qualità dell'aria. I dati raccolti e l'analisi è relativa alla zona _____.

L'interesse in questo argomento nasce dal fatto che pensiamo sia importante conoscere la situazione della qualità dell'aria e che grazie a delle previsioni sarebbe possibile controllare e forse "ridurre" l'inquinamento.

Un secondo obiettivo di questa ricerca, trovandoci in un periodo storico "particolare", ovvero la pandemia del Covid-19, è capire se quest'ultima oltre a cambiare le nostre abitudini abbia avuto o meno influenza sulla qualità dell'aria.

Data source(s)

I dati verranno presi giornalmente (4 misurazioni circa) dal sito Air Quality Programmatic APIs e collezionati in un dataframe tramite uno script python da noi sviluppato. Lo script si collega al sito tramite apposite API, fa richiesta della qualità dell'aria di ogni stazione _____ messa in lista ed inserisce le risposte in un dataframe che poi viene salvato in formato *.csv*. I dataframe di ogni osservazione verranno in seguito uniti.

Data collection

Di ogni stazione dell'aria raccoglieremo i seguenti dati: data, ora, località, indice AQI, monossido di carbonio CO, diossido di azoto NO_2, ozono O_3, anidride solforosa SO_2, particulate matter PM_{10}, particulate matter $PM_{2.5}$, umidità h, pressione p, temperatura t e velocità del vento w.

Il problema principale nella raccolta dati risiede nel fatto che le stazioni vengono aggiornate in orari diversi ed alcune vengono aggiornate soltanto una volta al giorno. Per questo motivo potremmo raccogliere più dati giornalieri di una stazione rispetto ad un'altra.

Model & Methods

I due modelli che vorremmo testare su i dati raccolti sono:

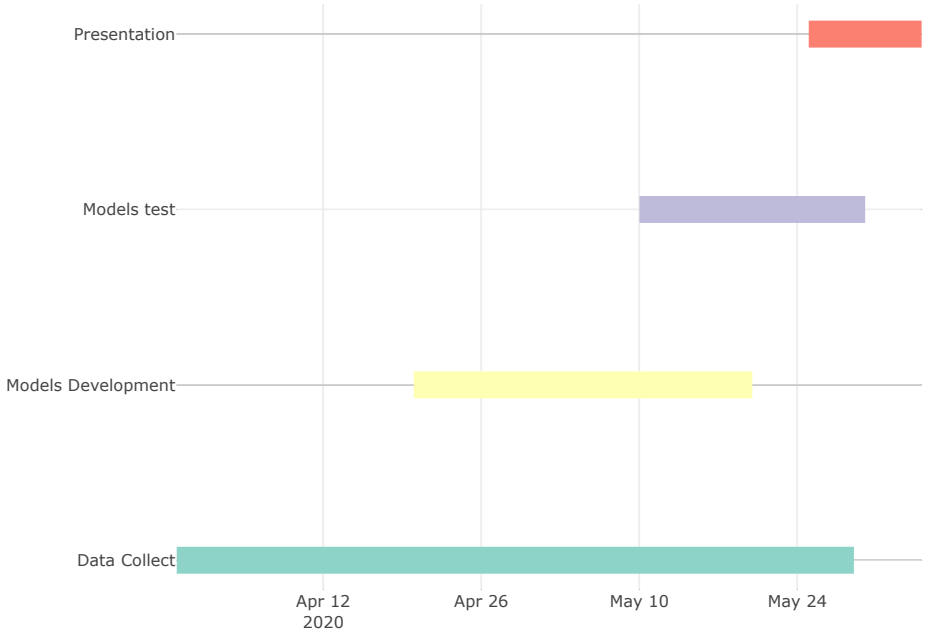
1. **Multiple Additive Regression Trees**: a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees (Wikipedia 2020b), (Karimian et al. 2019).
2. **ARIMA Model**: a generalization 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

In questo momento l'unico linguaggio di programmazione che abbiamo utilizzato è stato *Python*.

Per l'analisi dei dati continueremo ad utilizzare *Python* insieme ad *R*. Molto probabilmente faremo uso dei pacchetti gbm e 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.