CS361: Air Quality Index Forecasting

Harsh Katara Himanshi Gautam Dhruv Patel Ridhiman Kaur Dhindsa

Abstract

Forecasting urban air pollution is vital to combat its adverse impacts. Various machine learning methods will be utilized to predict air quality, with this study employing a variety of regression techniques. The aim is to forecast the Air Quality Index for major pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3. Evaluation criteria such as Mean Square Error, Mean Absolute Error, and R-squared error are to be used in determining the accuracy given by the various algorithms.

1. Introduction

1.1. Motivation

Air quality is crucial for public health and environmental management, impacting respiratory diseases and climate change. Accurate prediction of the Air Quality Index (AQI) is vital for interventions and policy-making. Worldwide, air pollution causes around 1.3 million deaths annually, with various harmful effects such as acid rain, global warming, and aerosol formation. The recent COVID-19 pandemic has highlighted the link between air pollution and higher death rates. Anticipating pollution fluctuations is urgent to mitigate its negative impacts. In India, the AQI ranges from 0-500, with eight major pollutants considered for calculation, including PM10, PM2.5, CO, O3, NO2, SO2, NH3, and Pb. AQI values correspond to different pollutant concentrations and health effects, emphasizing the importance of air quality evaluation in monitoring and controlling pollution.

Indian Context: Air pollution in India is a serious environmental issue. Of the 30 most polluted cities in the world, 21 were in India in 2019. As per a 2016 study, at least 140 million people in India breathe air that is 10 times or more over the WHO safe limit. 51% of the pollution is caused by industrial pollution, 27% by vehicles, 17% by crop burning and 5% by other sources. Air pollution contributes to the premature deaths of 2 million Indians every year. Emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm.

During autumn and spring months, extensive crop burning in agricultural fields, often chosen as a cost-effective alternative to mechanical tilling, emerges as a significant contributor to the proliferation of smoke, smog and particulate pollution. Furthermore, in the Indian context, seasonal variations are further compounded by notable spikes in air pollution attributed to the widespread combustion of firecrackers and heightened vehicular emissions, particularly during festivals like Diwali and Dussehra. These distinctive seasonal factors collectively exert a substantial influence on datasets related to air quality in India.

1.2. Target Problem

The primary goal of this project is to forecast AQI accurately. The AQI is a composite index that provides a simple way of describing the quality of air at a specific location and time based on several pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). The target problem involves predicting the AQI levels for the next few hours or days, considering various environmental factors.

1.3. Major Challenges

- *Data Availability and Quality:* Obtaining reliable and comprehensive air quality data for India might be challenging. Data may be sparse, inconsistent, or contain missing values, requiring careful preprocessing.

- *Feature Engineering:* Identifying relevant features such as meteorological data (temperature, humidity, wind speed), geographical factors, and historical AQI values and engineering them effectively for model input.

- *Model Selection:* Choosing an appropriate machine learning model that can effectively capture the complex relationships between various factors affecting air quality and provide accurate forecasts.

- *Interpretability:* Ensuring the interpretability of the model outputs so that stakeholders can understand the factors influencing air quality predictions and trust the forecasts.

2. Methods

To ensure that a diverse set of methods is used, we plan to use the following paradigms

Statistical Methods: We will leverage well-established statistical techniques to gain insights into our data. This includes employing Linear Regression and its sophisticated variants such as Lasso and Ridge regularization techniques. Additionally, we will harness the power of time series forecasting by utilizing renowned approaches like ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) models. These methods will enable us to capture the underlying trends and patterns in our data, facilitating accurate predictions and informed decision-making.

Ensemble Methods: Recognizing the strength in combining multiple models, we will employ ensemble methods to enhance the robustness and predictive performance of our analysis. Techniques such as Random Forests, Bagging, and Boosting will be utilized to aggregate the outputs of diverse models, thereby reducing variance and improving overall accuracy.

Kernel Methods: In pursuit of maximizing the discriminative capabilities of our models, we will explore the realm of kernel methods, with a focus on Support Vector Machines (SVMs). By employing SVMs with various kernel functions, such as linear, polynomial, and radial basis function (RBF), we aim to capture intricate nonlinear relationships within the data. These kernel methods offer flexibility and adaptability, allowing us to tailor our models to the specific characteristics of our dataset and potentially uncover hidden patterns that traditional linear methods may overlook.

3. Intended Experiments

In this project aimed at predicting Air Quality Index (AQI), several experiments can be conducted to explore different aspects of the problem and improve the predictive model. Here are some potential experiments:

1. **Feature Engineering:** Explore different viable datasets and experiment with feature sets within them such as weather data (temperature, humidity, wind speed etc.), historical AQI data, geographical factors (city, elevation, proximity to industrial areas or highways), time-related features (day of the week, hour of the day). The intent is to focus more on Indian cities and Guwahati in particular.

2. **Model Selection:** Experiment with different machine learning algorithms (as described in section 2) such as linear regression, decision trees, random forests, support vector machines and ensemble methods to determine which model performs best for AQI prediction given the dataset and features.

3. **Data Pre-processing:** Experiment with different data pre-processing techniques such as normalization, standardization, outlier detection, and missing value imputation to improve the quality of input data and model performance

4. **Error Analysis:** Experiment with techniques to analyse prediction errors and understand the factors contributing to inaccurate predictions. This could involve visualizing prediction errors, analysing feature importance, or conducting sensitivity analysis.

5. **Rolling Based Predictions:** Experiment with long term predictions for multiple time-stamps ahead based on the dataset. For eg. 14 days prediction, Monthly Predictions etc.

References

Bhattacharya, S. and Shahnawaz, S. Using machine learning to predict air quality index in new delhi, 2021. URL https://arxiv.org/abs/2112.05753.

Kaur, J., Singh, S., and Parmar, K. S. Forecasting of aqi (pm2. 5) for the three most polluted cities in india during covid-19 by hybrid daubechies discrete wavelet decomposition and autoregressive (db-dwd-arima) model. *Environmental Science and Pollution Research*, 30(45): 101035–101052, 2023. URL https://doi.org/10.1007/s11356-023-29501-w.

Kumar, A. and Goyal, P. Forecasting of daily air quality index in delhi. *Science of The Total Environment*, 409(24):5517–5523, 2011. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2011.08.069. URL https://www.sciencedirect.com/science/article/pii/S0048969711009661.