Predictive Analysis of PM2.5 Levels Using Machine Learning Techniques on Global Air Quality Data

Vittorio Balestrieri

Abstract

This project aims to analyze and predict PM2.5 air pollution levels using machine learning models, leveraging global air quality data accessed through the OpenAQ API. By focusing on PM2.5 due to its significant health implications, this study seeks to identify patterns, trends, and forecast future pollution levels, contributing to more informed environmental policies and public health strategies.

1 Introduction

Air pollution is a critical environmental issue with profound health impacts. Among various pollutants, particulate matter with a diameter of less than 2.5 micrometers (PM2.5) is particularly harmful. This project utilizes data science and machine learning techniques to analyze PM2.5 concentrations worldwide, aiming to uncover patterns and predict future levels. By integrating data from the OpenAQ platform, we demonstrate how technology can address pressing environmental challenges.

2 Methodology

2.1 Data Collection

Data was collected via the OpenAQ API, focusing on PM2.5 measurements across multiple countries. Due to API rate limits, strategies such as caching and request throttling were employed to efficiently gather a comprehensive dataset.

2.2 Data Preprocessing

The collected data underwent cleaning, including filtering for PM2.5 pollutants, handling missing values, and encoding temporal features.

2.3 Exploratory Data Analysis

Initial analysis included visualizing PM2.5 trends and distributions, which informed the selection of features for modeling.

2.4 Model Development

We employed machine learning models, starting with linear regression and advancing to more complex models like XGBoost, to forecast PM2.5 levels. Model performance was evaluated using metrics such as Mean Squared Error (MSE).

3 Results

Our findings highlight significant temporal and geographical variations in PM2.5 concentrations. The XGBoost model outperformed simpler models, demonstrating the potential of advanced machine learning techniques in predicting air quality.

4 Discussion

The predictive models developed in this study offer valuable insights for environmental monitoring and public health planning. Moreover, our methodology showcases the application of data science in environmental research, emphasizing the importance of leveraging technology for sustainability.

5 Conclusion

This project underscores the critical role of data science and machine learning in understanding and addressing environmental challenges. Future work will explore incorporating additional predictors, such as meteorological and traffic data, to enhance model accuracy.