Analyzing Air Quality: Insights into the Impact of Weather and Traffic in Big Cities

Related Work

Atreyo Das
das.at@northeastern.edu
Omer Seyfeddin Koc
koc.o@northeastern.edu
Shruti Suhas Kute
kute.s@northeastern.edu

Instructor: Dr. Fatema Nafa

Northeastern University, Boston MA
Master of Science in Data Science

Air quality prediction has been extensively researched, leveraging diverse machine learning models and datasets. This section explores related works and compares their methodologies with our approach, highlighting how our project advances the current state of research.

1 Air Quality Prediction Using Machine Learning (Raviteja et al.)

This study focuses on predicting air quality using machine learning models such as Random Forest, Decision Trees, KNN, and Gradient Boosting. Although the dataset used in their study is not explicitly specified, their work provides a relevant reference point as it employs some of the same models we plan to use in our project.

Raviteja et al. utilize Random Forest and Decision Tree models to predict air quality and analyze feature importance. Similarly, we plan to use Random Forest as a baseline model to explore the relationships between air quality and various influencing factors. However, their study does not consider advanced time-series forecasting methods. In contrast, we intend to integrate models like ARIMA and LSTM to analyze temporal patterns and predict future air quality trends, enabling a more forward-looking approach.

In terms of preprocessing, the paper outlines standard steps such as data cleaning and feature selection. While we will adopt similar preprocessing techniques, we also plan to include dynamic feature engineering to account for the interactions between weather, traffic, and pollutant data. This will allow us to gain deeper insights into the factors affecting air quality.

Furthermore, Raviteja et al. evaluate their models using metrics such as accuracy and precision, which are appropriate for classification tasks but may not fully capture the performance of regression or time-series models. In our project, we aim to use additional metrics like RMSE, MAE, and R^2 to evaluate the effectiveness of both predictive and temporal models.

For further details, the full text of Raviteja et al.'s paper can be accessed at the following link:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4878522

Contribution of Our Work

While our project builds on some of the methodologies used by Raviteja et al., such as machine learning models like Random Forest and Decision Trees, our approach is not entirely based on their study. Instead, we plan to extend their work by introducing advanced time-series forecasting techniques, such as ARIMA and LSTM, which were not explored in their research.

Additionally, unlike their focus on a single dataset, our project aims to analyze air quality by integrating multiple datasets, including traffic and weather data, alongside air quality measurements. This multi-faceted approach will allow us to conduct a more detailed and comprehensive analysis of the factors influencing air quality.

Although this paper serves as a valuable reference for parts of our methodology, we aim to expand upon their work to offer a broader perspective. By incorporating diverse data sources and advanced modeling techniques, we hope to provide more actionable and insightful analyses to help stakeholders better understand and mitigate urban air pollution.

2 Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities (Ameer et al.)

This study focuses on predicting air quality in smart cities using various regression techniques, including Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Multi-layer Perceptron Regression (MLP). The study evaluates these models based on their error rates and processing times, using datasets from five cities in China, covering the period from 2010 to 2015.

Ameer et al. highlight that Random Forest Regression outperformed other models in terms of accuracy and efficiency. Their findings suggest that Random Forest is particularly effective for datasets with significant historical data, offering the lowest error rates (MAE and RMSE) and reasonable processing times compared to other models. However, the study primarily focuses on static datasets and does not incorporate time-series forecasting techniques or multiple contributing factors such as traffic or weather data, which limits its scope for dynamic urban air quality prediction.

In terms of preprocessing, the study includes standard data cleaning and correlation analysis but lacks advanced feature engineering to capture interactions between multiple environmental variables. The evaluation is based solely on pollutant concentrations (e.g., PM2.5) and meteorological variables, without integrating additional real-time urban data sources.

For further details, the full text of Ameer et al.'s paper can be accessed at the following link:
https://doi.org/10.1109/ACCESS.2019.2925082

Contribution of Our Work

While our project draws inspiration from the methodology employed by Ameer et al., it aims to address several of the limitations in their study. First, we plan to extend their work by incorporating advanced time-series forecasting models such as ARIMA and LSTM, which are more suitable for predicting future trends in air quality. Second, we will integrate multiple datasets, including traffic and weather data, alongside pollutant measurements, to provide a more holistic analysis of air quality in urban areas.

Moreover, unlike Ameer et al., who conducted their study using datasets from cities in China, we will focus on major urban areas in the United States. By utilizing publicly available data sources like NYC Open Data, we aim to conduct a more detailed and region-specific analysis. Additionally, we will explore their findings on feature engineering and apply these techniques in our work to identify and model complex interactions between variables.

Our project will also go beyond static datasets by incorporating dynamic and interactive data analysis through dashboards, enabling real-time decision-making for stakeholders. By combining diverse data sources and advanced modeling techniques, we aim to offer actionable insights that support urban planners and policymakers in mitigating air pollution.