Predicting Air Pollution Levels in India Using Multivariate Regression Analysis

Abstract

Air pollution remains one of India's most pressing environmental challenges, significantly impacting public health, the economy, and climate change. This project aims to predict air pollution levels in India by developing a multivariate regression model using various predictors such as location, season, weather data, temperature, and other relevant environmental and socio-economic factors. By leveraging data mining techniques and analyzing existing literature, the study seeks to identify key determinants of air pollution and provide insights that could inform policy decisions and mitigation strategies.

Contents

1 Introduction

Air pollution in India has reached alarming levels, with several cities consistently ranking among the most polluted globally. This pollution is generated by various chemical pollutants, such as carbon monoxide (CO), ozone (O_3), nitrogen oxides (NO_x), sulfur oxides (SO_x), and particulate matter (PM_{10} and $PM_{2.5}$). The adverse effects of air pollution extend beyond environmental degradation, affecting human health, economic productivity, and contributing to climate change. According to the World Health Organization, air pollution is a leading cause of premature deaths worldwide, with India accounting for a significant share [1].

Understanding the factors contributing to air pollution is crucial for developing effective mitigation strategies. This project aims to build a predictive model for air pollution levels by analyzing data from eight major Indian cities. At this stage, the most suitable modeling techniques and the specific dependent variables have yet to be determined. It is not yet clear whether the developed model will predict pollution levels on a daily or hourly basis. This decision will depend on factors such as data availability, complexity, and the desired balance between model accuracy and manageability, making it a key consideration for the study.

2 Data Collection and Initial Analysis

2.1 Data Sources

This project utilizes two primary sources of data to analyze air pollution levels across eight major Indian cities:

- **Air Quality Data in India**: Available at Kaggle. This dataset provides hourly measurements of various air pollutants and particulate matter. The data is collected from multiple weather stations located within each of the eight major cities. Recognizing that larger cities may have several monitoring stations to capture spatial variability in air quality, we aggregate the pollutant levels by averaging the measurements from all stations within a city for each hour. This averaging process ensures that the data represents the overall air quality of the city rather than isolated monitoring points.

- **Historical Weather Data for Indian Cities**: Available at Kaggle. This dataset includes hourly weather-related features for the same set of cities, encompassing over 20 variables such as precipitation (mm), wind speed, temperature, humidity, and other meteorological parameters.

Merging these two datasets is a strategic choice for several reasons. Firstly, it allows for a comprehensive analysis of the relationship between air pollution levels and various weather conditions. By integrating pollutant concentrations with corresponding meteorological data, we can better understand how factors like temperature, wind speed, and precipitation influence air quality. This holistic approach enhances the model's ability to capture the multifaceted nature of air pollution dynamics.

Furthermore, we extended the merged dataset by incorporating precise geographical coordinates for each of the eight cities analyzed. Using the Google Maps API, we retrieved

the exact latitude and longitude for each city, adding these as additional predictors in our model. Including geographic coordinates is expected to account for spatial dependencies and regional differences in pollution patterns, thereby potentially improving the model's predictive accuracy.

In summary, the final dataset comprises hourly air quality measurements, detailed historical weather data, and geographical information for eight major Indian cities. This integrated dataset provides a robust foundation for developing a predictive model aimed at forecasting daily air pollution levels, balancing data comprehensiveness with manageability.

2.2 Initial Findings

TODO

3 Proposed Methodology for Prediction

3.1 Regression Model Selection

- **Multivariate Linear Regression**: To model the relationship between air pollution levels and continuous predictors.
- **Support Vector Regression (SVR)**: For handling non-linear relationships in the data.
- **Random Forest Regression**: To capture complex interactions between variables and improve prediction accuracy.

3.2 Variables Considered

Independent Variables (Regressors):

- **Location**: Latitude, longitude, and elevation.
- **Season**: Categorical variable representing different seasons.
- **Temperature**: Daily average temperature.
- **Humidity**: Atmospheric moisture content.
- **Wind Speed**: Affects dispersion of pollutants.
- **Industrial Activity**: Proximity to industrial zones.
- **Traffic Density**: Number of vehicles in the area.
- **Population Density**: Human activities contributing to emissions.

Dependent Variable:

- **Air Pollution Level**: Concentration of pollutants like PM2.5 and PM10.

3.3 Data Preprocessing

- **Handling Missing Values**: Imputation techniques or exclusion based on data availability.
- **Normalization**: Scaling variables to ensure uniformity.
- **Encoding Categorical Variables**: Converting seasons and other categorical data into numerical format using one-hot encoding.

3.4 Model Training and Validation

- **Training Set**: 70% of the data for training the model.
- **Validation Set**: 15% for tuning hyperparameters.
- **Test Set**: 15% for evaluating model performance.

3.5 Evaluation Metrics

- **Mean Squared Error (MSE)**: Measures the average squared difference between observed and predicted values.
- **R-squared ($\mathbf{R^2}$)**: Indicates the proportion of variance explained by the model.
- **Mean Absolute Error (MAE)**: Provides the average magnitude of errors in predictions.

4 Expected Outcomes

- **Identification of Key Predictors**: Understanding which variables significantly impact air pollution levels.
- **Predictive Model**: A reliable model that can forecast air pollution levels based on the given regressors.
- **Policy Implications**: Insights that can help in formulating targeted interventions to reduce pollution.

5 Conclusion

TODO

References

[1] S. Dey and [Other Authors]. "Ambient Air Pollution and Daily Mortality in Ten Cities of India: A Causal Modelling Study". In: *The Lancet Planetary Health* 4.7 (2020), e287–e298. DOI: 10.1016/S2542-5196(24)00114-1. URL: https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(24)00114-1/fulltext.