An Analysis of Air Quality in Delhi: Assessing the Impact of PM2.5 and Forecasting Future Trends for India

Ayushi Dey
ayushi.dey2022@vitstudent.ac.in

Vinit D. Bangera
vinit.bangera2022@vitstudent.ac.in

Guide:
Dr. Jitendra Kumar
Department of Mathematics
Vellore Institute Of Technology
jitendra.kumar@vit.ac.in

Abstract: ***India ranks eighth on the list of nations with the worst air quality index, and twelve of the fifteen most polluted cities in Central and South Asia are located within its borders. Sixty percent of the cities analyzed in India had annual $PM_{2.5}$ levels that were at least seven times higher than the WHO's guidelines. Delhi's Air Quality Index (AQI) is the most severe of all states and union territories in the country, with $PM_{2.5}$ levels that are nearly 100 times the WHO safe limit. The purpose of this study is to investigate and analyze Delhi's air quality from January 2021 to May 2023 and identify data trends. In addition, $PM_{2.5}$ levels in the air have been forecasted for India.***

Keywords: *Air Pollution, $PM_{2.5}$, India, Delhi, Pollutants, Air Quality, LSTM, Random Forest, SVM, SARIMA*

I. Introduction

Delhi is one of the most polluted cities on earth. IQ Air's recent report on air quality rates Delhi as the most polluted capital city among 106 nations based on $PM_{2.5}$ concentration. According to the World Health Organisation, Delhi is the sixth most polluted of India's 13 main cities. In October of 2019, $PM_{2.5}$ concentrations in Delhi reached 440 g/m^3, which is twelve times the limit recommended by the USA. This research aims to analyze the recent air quality of Delhi beginning in January 2021 and ending in May 2023, and to develop a time series model for India using air quality data from 2017 to 2022 to forecast the $PM_{2.5}$ level in the air. Our research is pertinent to the residents of Delhi, the state and federal governments, and the health department. By forecasting the $PM_{2.5}$ level in the air, we can strive to control it as necessary and help provide cleaner, safer air to breathe.

II. Literature Review

Nilesh N. Maltare et. al. (2023)[1] studied the air quality index (AQI) of Ahmedabad, Gujarat and developed various machine learning models like SARIMA, SVM and LSTM to predict the AQI. It was found that on numerous evaluation metrics like R2 score and RMSE, SVM with RBF kernel outperformed other SVM model modifications and models. Dutta A et. al. (2017)[2] investigated Delhi's air quality and the relationship between air quality and respiratory diseases. Bhatti et. al. (2021)[3] studied the air quality of Lahore which revealed that the seasonal particulate matter ($PM_{2.5}$ and PM_{10}) in Lahore exceeds Pakistan's National Environmental Quality Standards (NEQS). Studies of correlation indicate a positive relationship between particulate matter and other mass-concentration particles such as ozone (O_3), nitrogen oxide (NO), and sulphur dioxide (SO_2). Higher CO/NO ratios indicate that mobile sources are one of the primary contributors to this increase in NO.

III. Data

A. Data Description

This study is conducting for the city of Delhi (28.7041° N, 77.1025° E) from India. The data is collected from various stations of Delhi through Delhi Pollution Control Committee. The dataset is collected from the year 2021 to May, 2023, observed from 11 different monitoring stations. Forecasting of $PM_{2.5}$ has been done based on a dataset collected from the year 2017 to 2022 from the Central Pollution Control Board.

B. Outlier Analysis

The z-score method was used to analyze outliers. The z-score represents how many standard deviations away the actual value is from the mean. By setting a limiting value for the z-score, we can

identify and label data points as anomalies in the overall context.

$$z = \frac{x-\mu}{\sigma} \quad -(i)$$

here, x is the data point, σ is the standard deviation and μ is the mean of population.

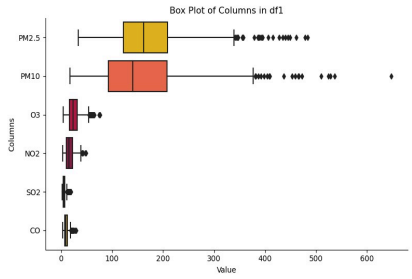

Fig. 1: Graphical representation of outliers using boxplot

C. *Handling Missing Value*

There were not many missing values present in the dataset of Delhi. Hence, deleting rows was a suitable solution to handling the missing values.

IV. Methodology

A. *Exploratory Data Analysis*

Exploratory data analysis (EDA) typically occurs early in the analytic process and pertains to initial analyses and conclusions drawn from data sets. Typically, box plots are used to visually depict outliers. Using correlation matrices, correlation coefficients between various variables are represented. A bar graph is used to graphically represent data. It employs bars of varying heights to represent value. A graph with multiple bars depicts two or more interconnected data sets (many bar diagrams facilitate the comparison of numerous phenomena). Line graphs are employed to illustrate trends and patterns in data over time.

B. *Descriptive Analysis*

The coefficient of variation (C.V.) of each pollutant is calculated and consistency ranking is done; lower the coefficient of variation, lower is the rank.

$$C.V. = \frac{Mean}{Standard\ Deviation} \times 100 \quad -(ii)$$

C. *Modeling*

SVM is a supervised algorithm for ml used for classification and regression tasks. It is a supervised learning model that analyzes data and identifies the optimal hyperplane that separates classes in the feature space.

SARIMA (Seasonal ARIMA) is beneficial when a time series exhibits seasonal variation (Dubey et al., 2021). The SARIMA may be portrayed as

$$\phi_p(B)\Phi_P(B^S)(1-B)^D(1-B^S)^D Z_t = \theta_q(B)\Theta_Q(B^S)$$
$$a_t \quad -(iii)$$

LSTM is a recurrent neural network architecture used extensively for sequential data analysis, including time-series forecasting, NLP and speech recognition.

Random Forest Algorithm is a supervised machine learning algorithm utilized extensively in Machine Learning to solve Classification and Regression problems. Random Forest is composed of multiple decision trees which are applied to various subsets of a given dataset and the average of it is used to improve the accuracy of that model.

V. Results & Discussions

A. *Exploratory Data Analysis*

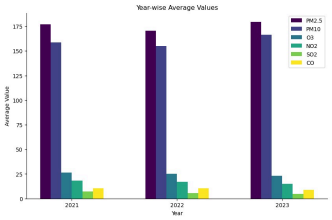

Fig. 2: Correlation Matrix of Pollutants

Fig. 2, shows that $PM_{2.5,}$ PM_{10} had the highest annual average from 2021 to May 2023. Compared to all other pollutants, the annual averages of $PM_{2.5}$ and PM_{10} are significantly higher, indicating that particulate matter is the primary pollutant in Delhi's air.

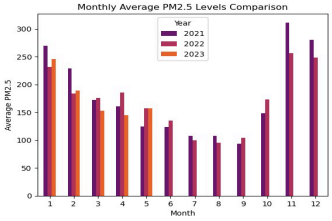

Fig. 3: Monthly Average $PM_{2.5}$ Levels Comparison

Further focusing on $PM_{2.5}$ and from fig. 3, it was discovered that the monthly averages of $PM_{2.5}$ in 2021 was significantly higher as compared to those as of May this year.

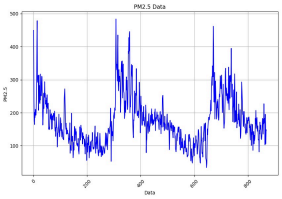

Fig. 4: $PM_{2.5}$ Data

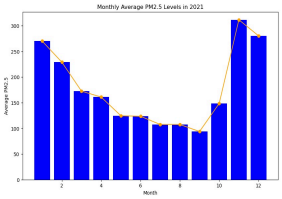

Fig. 5: Monthly Average of $PM_{2.5}$ levels in 2021

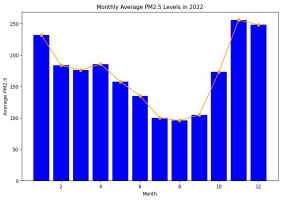

Fig. 6: Monthly Average of $PM_{2.5}$ Levels in 2022

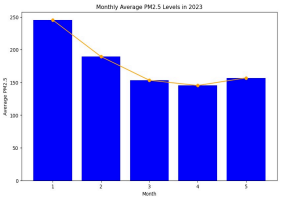

Fig. 7: Monthly Average of $PM_{2.5}$ Levels in 2023

From fig. 5 and 6, in November of both 2021 and 2022, the monthly average attained its highest point for the year, closely followed by December. This may indicate that $PM_{2.5}$ concentrations tend to be higher at lower temperatures.

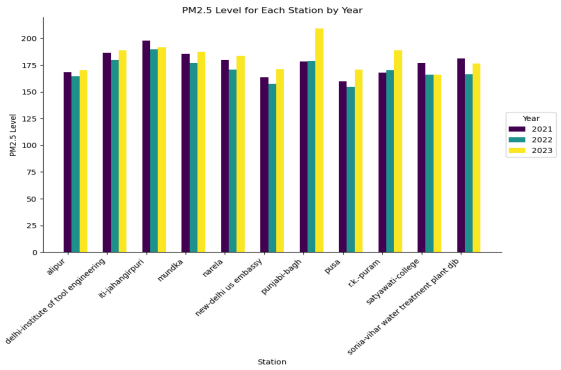

Fig. 8: Average PM2.5 Level of Each Station by Year

For this study, 11 stations have been taken into consideration and it was found that $PM_{2.5}$ level was comparatively lower for all the stations in the year 2022, as shown in fig. 8.

B. *Descriptive Analysis*

From Table 1, it can be inferred that PM_{10} has the highest coefficient of variation while CO has the lowest coefficient of variation.

Table 1: Coefficient of Variation Table

Rank	Pollutant	Coefficient Of Variation
1	PM_{10}	54.8916
2	SO_2	49.7733
3	O_3	45.4091
4	NO_2	44.4266
5	$PM_{2.5}$	42.3557
6	CO	36.1768

C. *Modeling*

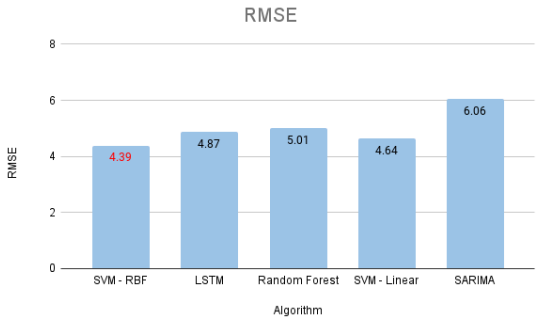

Fig. 9: Comparative Analysis of RMSE Scores of Various Models

VI. Conclusion

This research focuses on creating a reliable model to predict the $PM_{2.5}$ in Delhi using the CPCB air quality dataset. The study employs various preprocessing techniques, including outlier removal, feature selection, and handling missing values, to improve the representation of the data. Machine learning models, namely LSTM, SARIMA, SVM with different kernel functions, and Random Forest with hyperparameter tuning, were evaluated to develop an accurate predictive model. Among these models, SVM with an RBF kernel outperformed the others in predicting $PM_{2.5}$ data for India, based on multiple evaluation metrics like RMSE. The future scope of this research involves assessing the proposed model's effectiveness in predicting AQI data in different geographical regions.

VII. References

[1]Nilesh N. Maltare, Safvan Vahora, Air Quality Index prediction using machine learning for Ahmedabad city, Digital Chemical Engineering, Volume 7 (2023)

[2]Dutta A, Jinsart W. Air pollution in Delhi, India: It's status and association with respiratory diseases. PLoS One. 2022 Sep 20;17(9)

[3]Bhatti, Uzair & Yuhuan, Yan & Ming-Quan, Zhou & Ali, Sajid & Hussain, Aamir & Qing-Song, Huo & Yu, Zhaoyuan & Yuan, Linwang. (2021). Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM 2.5): An SARIMA and Factor Analysis Approach.

[4]Srivastava, S., Kumar, A., Bauddh, K. *et al.* 21-Day Lockdown in India Dramatically Reduced Air Pollution Indices in Lucknow and New Delhi, India. *Bull Environ Contam Toxicol* 105, 9–17 (2020)