DELHI AIR QUALITY INDEX (AQI) MONITORING SYSTEM

Apurva M. Bhavsar[1], Shubham S. Khairnar[2], Dhruv R. Patil[3], Atharva P. Patil[4], Saurabh A. Shinde[5]

[1]Asst. Professor, Dept. of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering and Research Centre, Nashik, Maharashtra, India.

[2,3,4,5]Student, Dept. of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering and Research Centre, Nashik, Maharashtra, India.

ABSTRACT

The timely and accurately forecasting of air quality is one of the major determinants of urban sustainability. In this study, the PM2.5 concentration in Delhi is predicted using machine learning techniques from environmental and pollutants data. The models evaluated in this study were Random Forest and XGBoost, both capable of modelling complex and nonlinear phenomena. Data were collected from open repositories, pre-processed through imputation and normalization, and the appropriate features such as NO, NO2, CO, SO2, O3, PM10, NH3 were selected and used for training the models. The results indicate that XGBoost is slightly better than Random Forest regarding predictive capability. This would find utility in issuing public health-related advisories and urban planning.

Keywords: Air Quality, PM2.5, Random Forest, XGBoost, Prediction, Machine Learning

1. INTRODUCTION

In recent decades, energy consumption has expanded quickly due to accelerated urbanisation and industrialization, resulting in severe air quality problems. Recently, air pollution incidents have become more frequent in China, posing a very severe threat to inhabitants' health. Air pollution disperses hazardous substances in the form of particulate matter and gaseous pollutants in the surrounding atmosphere that are harmful to the environment. Environmental air pollutants includes particulate, tropospheric ozone, carbon dioxide, sulphur dioxide, etc. Consequently, air quality should severely upset the ecosystem balance, producing greenhouse gases, acid rain, ozone layer depletion, and many more. However, numerous studies have been conducted to show that long-term air pollution has a wide range of negative health effects, including respiratory and cardiovascular problems.

To predict small changes in air pollution concentrations are very crucial and difficult, especially in case of limited data inputs and high parameter variability. In such situations, high computational power is required, which can be achieved through machine learning and deep learning approaches. Machine learning is a branch of artificial intelligence that lets a machine learn from different types of data/information. Deep learning is another kind of learning that provides the capability to learn from huge data. Other than some applications in air pollution too.

Since the AQI follows a periodic pattern, the deep learning models could be used for accurate prediction of AQI values. In India, rapid urbanization and industrial growth have degraded air quality in many places, especially big cities like Delhi where PM2.5 pollution has become a salient risk factor for the respiratory and cardiovascular health of the people. Therefore, predicting air quality levels is necessary for intervention in environmental monitoring, as well as for public health interventions.

Traditional forecasting techniques largely ignore the complexity that relates to air pollution. The strength of machine learning in modeling non-linear relationships provides a more appealing alternative. This study harnesses environmental data and machine learning models to predict PM2.5 levels with the explicit aim of enforcing timely decision-making to minimize anticipated exposure risks.

2. METHODOLOGY

The research objective was the design of a system that predicts PM2.5 concentration levels in Delhi through a machine-learning regression model. The entire work has data pipeline as follows: data acquisition, preprocessing data, data transformation, model selection, model training and performance evaluation of models. The front end is interactive and was designed in Streamlit for easy practical deployment and usability.

2.1 Data Collection

The data collected for the present study was drawn from publicly obtainable air quality monitoring components and it is taken from kaggle. It contained historical data on diverse pollutants and environmental conditions as recorded in Delhi. The prominent features contained:

Concentration of pollutants: NO, NO_2, SO_2, CO, O_3, PM10, NH_3

Meteorological variables: Temperature, Humidity. Target variable: PM2.5 concentration

2.2 Data Preprocessing

2.2.1 Handling Null Values

Missing values in environmental datasets can arise from sensor malfunctions and errors during data transmission. The various statistical methods can be employed to fill up the numerical missing data and to resolve these discrepancies. Such methods are analyzing the distribution of each feature to decide what mean or median should be taken for imputation considering the skewness of the data so that the integrity of the data is maintained along with accuracy to that of the original information. This imparts reliability of the dataset for further analysis and interpretation. This technique further employs the statistical imputation methods to minimize the effect of lack of data on results and help make more rigorous conclusions. Overall, careful missing data management for environmental datasets helps to optimize the analyses validity and quality for more accurate insights into the environmental phenomenon.

2.2.2 Label Encoding.

Most parameters in almost any dataset are numerical, yet some categorical parameters like station name and date-time fields seem more common. It is occasionally called Label Encoding for the sake of usability and interpretability. This process involves the application of Scikit-learn facilities, such as Label Encoder, to convert non-numeric or categorical variables into a numeric representation in order to develop a framework in which these categorical values have a consistent numeric equivalence but, at the same time, maintain their original label semantics. So, this utility serves to convert such categories from the usual categorical labels into numeric forms that can be used with respect to such categories for any analytical purposes. Converting categorical attributes to numerical encoding prepares them for calculation and modeling of accurate analytical evaluation. Thus, in this sense, different data of any single dataset act altogether for a push of the data because of its value at different analytical instances.

2.3 Feature Selection and Scaling

The model efficiency was upgraded by performing an elaborate bivariate feature correlation analysis using a visually explanatory heatmap. The heatmap allowed for conveniently pinpointing and underlining particular feature coordinates that had a high correlation with the target variable (PM2.5) identified as primary features. Further, the heatmap was influential in understanding the relationship structure between the variables with the prime aim of selecting salient features that could be useful for predictive modeling. One observation of interest was that the feature under consideration did not seem to require scaling in this modeling exercise. An interesting aside is that, in future model extensions, feature scaling may be a good option to include, especially in the case of distance-based models requiring normalization. In consideration of this, future model versions would probably have an advantage from the feature scaling being considered for improving performance especially in cases where distance-based algorithms would be involved. This way of including refinements and improvements will provide more opportunity for fine-tuning and optimizing any future models in the area of PM2.5 forecasting and analysis to handle the much more complex data relationships and increase the prediction accuracy.

2.4 Regression Models

The prediction task was carried out by means of two models:

- Linear regression: A baseline model assumes a linear relationship between input variables and target variables.
- XGBoost regressor: A powerful gradient boosting algorithm that emphasizes high performance and accuracy. It incorporates regularization, tree pruning, and handling of missing values.

Both models were trained using an 80:20 train-test split. The feature importance was then visualized to identify the major influencers for PM2.5 concentrations.

2.5 Streamlit Deployment

With a view to enhancing accessibility and user engagement, the prediction systems in their full form were decided to be put up as a user-friendly web application based on the Streamlit platform. This choice facilitates not only the accessibility of the system to the user but also promotes higher levels of interactivity and real-time data analysis. Under this application environment, users can upload their datasets in an easy manner in the CSV format for data to be analyzed. It is further designed to allow users a choice between various models including the two popular ones: Linear Regression and XGBoost, thus providing a flexible and customized avenue for data analysis giving users a chance to choose the most suitable model with regard to their datasets and analytical goal. The objective is to put all these features into a web-based application for a user-friendly and interesting environment to facilitate the insightful and efficient exploration of their data. This setup enabled seamless interaction with the system by users and the meaningful application of the results generated by their selected model to support decision-making to enhance their overall experience and the prediction system's usefulness.

3. MODELING AND ANALYSIS

The present study was aimed at modeling and analyzing the various machine learning algorithms and observing their behavior with real data. The focus was on determining the effectiveness of various regression techniques in predicting PM2.5 particulate concentration using historical air quality and meteorological data collected from the city of Delhi.

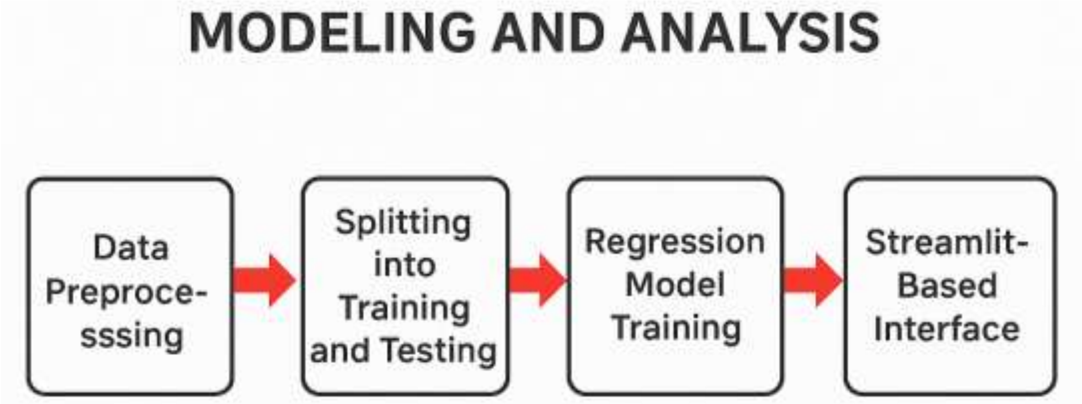

MODELING AND ANALYSIS
Data Preproce-sssing
Splitting into Training and Testing
Regression Model Training
Streamlit-Based Interface

Figure 1: Modelling And Analysis Procedure.

3.1 Model Implementation Environment

In performing all modeling as well as testing, one uses the features of the famous language called Python. The major libraries used in the present case are:

- Pandas and NumPy for manipulation of data.
- Scikit-learn for regression model building, preprocessing, and evaluation.
- Xgboost for building the model by realizing the concept of gradient boosting for classification and regression.
- Matplotlib and Seaborn for visualizations.
- Streamlit for setting up the user interface of the web-based prediction tool.

This design of environment was meant for modularized experimentation over all possible machine learning models and preprocessing pipelines.

3.2 Regression Model Training

The system created above on the gathered information through preprocessing after which it was made a division into training (80%) and testing (20%) subsets so as to carry out unknown data evaluations on the models to avoid overfitting. The regression model used which is as follows. These include the following;

3.2.1 Linear Regression

This serves as the baseline for the models under study. It uses the data to fit a linear function using the least squares criterion minimizing square residuals between observed and predicted values for PM2.5. Although being very simple, by giving some information from the result, the linearity of the feature-target relationship is usually apparent while also making it computationally cheap.

Model Behavior

- There were three characteristic features of the model behavior as observed:
- Reasonable performed on features with direct dependence or proportional influence on PM2.5
- Captures minimally non-linear or mutually interactive effects between pollutants

3.2.2 XG Boost Regressor

The current model is so chosen due to its superiority in handling the non-linearity besides other regularization features and also shows robustness in overfitting. The model constructs an ensemble of decision trees where each successive tree corrects the error of the preceding one.

Hyper parameters used

The Number of Estimators – 100. Learning Rate – 0.1 Maximum Depth – 6

Model Behavior

- With slight dependency between the levels of all pollutants, it has a good prospect of capturing something even higher.
- Compared to Linear Regression, it deals with missing data and skewed feature distributions more effectively.

3.2.3 Random Forest

Random Forest is an ensemble learning method that employs internodes multiple decision trees for better prediction accuracy, reducing overfitting. In the Delhi AQI Prediction System context, Random Forest becomes an option for either Linear Regression or XGBoost as a non-linear regression model with robustness.

Model Behavior

- Random Forest would better prove to be more in terms of effective completion involving the R² Score and RMSE but still proves to fall slightly short of being better than XGBoost.
- Speedy compared to XGBoost when training, but has no optimization options for real-time or even very large-scale deployment because of its memory consumption.

3.3 Visualization and Interpretation

For the purpose of finding an understanding of model behavior and data structure, different visualizing methods were involved during the analysis:

3.3.1 Heatmap of Correlation

A heat map concerning the correlation between certain environmental factors in relation to PM2.5 versus other pollutants has been constructed on the Pearson correlation coefficient. This heatmap will help forecast the extent of a linear relationship that exists between PM2.5 and other pollutants, namely between PM2.5 and NO2, PM10, as well as CO. It will particularly show how strongly these strong positive correlations are with each other. These were the variables that thus have been seen to affect PM2.5 levels meaningfully and have therefore been included in the above-mentioned model. Subjects are able to see what some hot spots are in terms of contributing factors to PM2.5 concentration with respect to interrelatedness data from the heatmap. Overall, this will serve as a good measure of how the variables come together in combining effects relative to PM2.5 concentrations. This visualization tool provides an overview in a truly panoramic view of how these critical environmental factors are related to each other concerning their total effects while also highlighting possible patterns or trends that would merit additional analysis. Definitionally, the correlation heatmap is an essential analytical tool for establishing the complexity concerning PM2.5 and related pollutants. Therefore, it will help in shaping more effective pollution control strategies and policies.

3.3.2 Actual vs Predicted Plot

One very effective visual method of demonstrating predicted accuracy could be the graph plotting actual concentrations of PM2.5 against those forecasted by the model. Here is where the actual power of the performance of the model shows: the closer they are, the more output of our model becomes a fairly reliable, precise indicator of its own self. And in this respect, the XGBoost model has invariably borne a handsome similarity with almost no bias, thus explaining the great propensity for its accuracy in prediction. For this tenacity is a testimony to its robustness, as well as a testimony to its model's ability to produce the right results over and over again. Therefore, it has the capability to show that kind of graph has great strength in representing the capability and efficiency of a model, rightly predicting PM2.5 values. Thus, it can further substantiate the level of correspondence shown in the graph between the predicted and the actual value of understanding the performance of a model and thus enhances the credibility associated with it to be used for high accuracy in PM2.5 predictions. Thus, underscoring the potency of using visual aids to authenticate and understand the performance of models like XGBoost in applications regarding monitoring and assessment of the environment.

3.3.3 Feature Importance (XGBoost)

The importance of XGBoost feature selection lies in the clarity with which it states the influence of variables on prediction making with the heart of input variable contribution analysis: Some features are said to be important. That is, of the variables which are considered important for making predictions, NO, NO_2, PM10, CO, NH_3, $O_{3,}$ SO_2 and Temperature are key. Their being key in understanding whether and how air quality predictions are dependent on other variables will also help researchers and analysts to work on improving model performance and prediction accuracy. Besides that, knowing which environmental variables are important for air quality variation is valuable information for developing appropriate pollution control and mitigation strategies. Accordingly, focusing on these core features will allow the stakeholders to make decisions and implement strategies for improving air quality for the well-being of the public. To end here, it is interlinked with recognizing and prioritizing these variables in advancing the understanding of air quality dynamics and thus providing pathways to data-driven solutions for a clean and healthy environment.

3.4 Streamlit Based Interface

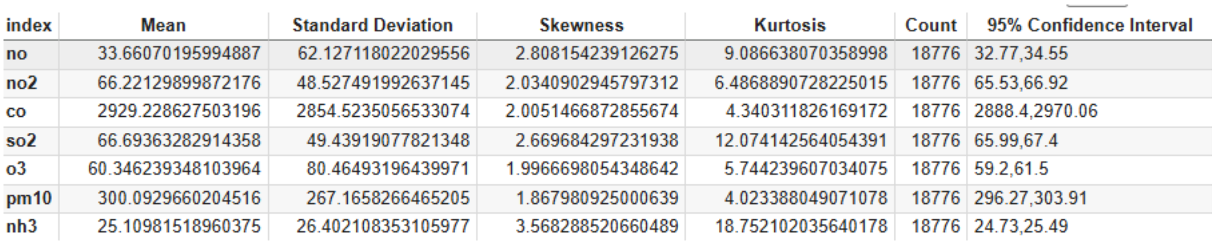
The Streamlit interface becomes an excellent human-machine interaction model. Anyone who wants to engage in AQI predictions without computer programming could use this human-centric web application. Thus, the Streamlit philosophy allows interaction between AQI prediction models, data points, and visualizations. The fun part is that it enables user interaction-working on inputs and giving feedback on a live basis. Programming is perhaps the biggest bottleneck that cannot be tolerated here. Through this simple avenue, the stunning predictions of the AQI would be all there for visualization and analysis, greatly enhancing user engagement and understanding. The fusion of data science and human interaction through the Streamlit interface has shown technology simplifying in real sense for a heterogeneous lot of users.

4. RESULTS AND DISCUSSION

4.1 Descriptive Statistical Analysis

The descriptive statistics with regard to the pollutants in the Delhi AQI dataset have shown high variability yet skewedness for the gases. Considering mean concentration, CO had the highest concentration (2929.23 µg/m³) followed by PM10 (300.01 µg/m³) and NO₂ showing a concentration of 66.22 µg/m³. The other remaining pollutants are positively skewed and highly kurtosed, which indicates outliers and heavy-tailed distributions. NH₃ and SO₂ are highly positive in skewness and kurtosis, indicating infrequent extreme values in similar lines as the above pollutants. Since there will always be 18,776 entries for all pollutants, very statistically sound analyses can be performed, generating really narrow 95% confidence intervals that translate into reliable mean estimates for each pollutant.

Table 1: Descriptive statistics of pollutants



index	Mean	Standard Deviation	Skewness	Kurtosis	Count	95% Confidence Interval
no	33.66070195994887	62.127118022029556	2.808154239126275	9.086638070358998	18776	32.77,34.55
no2	66.22129899872176	48.527491992637145	2.0340902945797312	6.4868890728225015	18776	65.53,66.92
co	2929.228627503196	2854.5235056533074	2.0051466872855674	4.340311826169172	18776	2888.4,2970.06
so2	66.69363282914358	49.43919077821348	2.669684297231938	12.074142564054391	18776	65.99,67.4
o3	60.346239348103964	80.46493196439971	1.9966698054348642	5.744239607034075	18776	59.2,61.5
pm10	300.0929660204516	267.1658266465205	1.867980925000639	4.023388049071078	18776	296.27,303.91
nh3	25.10981518960375	26.402108353105977	3.568288520660489	18.752102035640178	18776	24.73,25.49

4.2 PM 2.5 Time Series Analysis

The timeline PM2.5 graph is thus indicating the time-dependent fluctuations of fine particulate matter in Delhi. The graph shows extensive coloration on interval spikes above 1000 µg/m³, indicating episodes of very severe air pollution. All of the above suggest a general trend that should indicate a very high level of pollution in an apparently ongoing manner but with transient variations. Therefore, it is presumed that air quality has been seriously compromised for quite some time. Peaks certainly clustering closely point to the severity and frequency of pollution events, presumed to be linked with traffic, industrial activities, and seasonal changes, which underlines the need for ongoing air quality monitoring and mitigation programs.

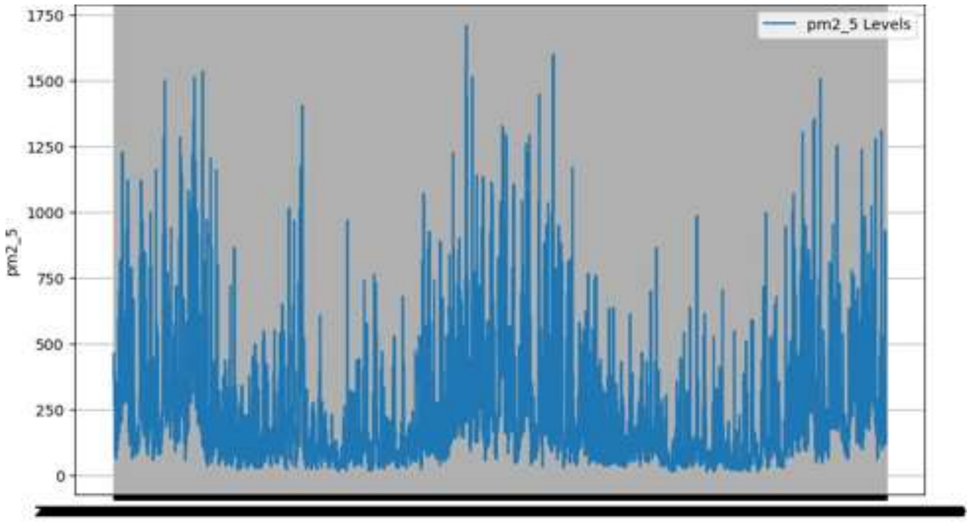


Figure 2: PM 2.5 Time Series

4.3 Correlation Heatmap of Pollutants

A correlation heatmap shows how the concentrations of different air pollutants relate to each other in the Delhi AQI data. Most of those pollutants have a very high level of correlation: PM2.5 and PM10 (0.99). CO, NO, and NO2 are also very much correlated. In contrast, O3 has weak or even negative correlations with most other elements because of its photochemical nature and different formation dynamics. This heatmap is helpful in getting the pollutant pair with similar behavior rates that would, in turn, lead to the development of better prediction models and air quality management strategies.

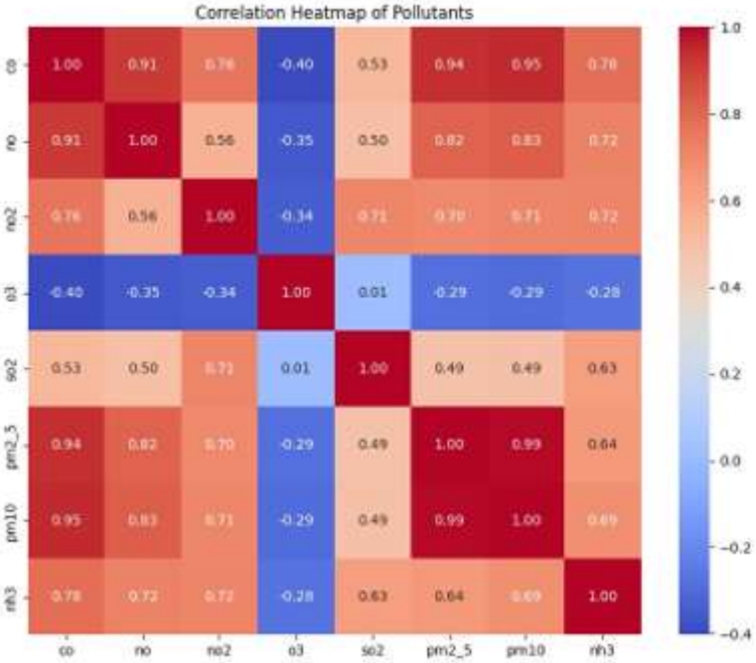


Figure 3: Correlation Heatmap of Pollutants

4.4 Feature Importance of Coefficient of Pollutants

The feature importance plot based on coefficients for specific gases provides a clear indication of the contribution by each pollutant to the overall prediction of AQI in a linear modeling framework. On PM10, it clearly has shown a significant positive coefficient, denoting its primary importance for deteriorating air quality. Closely followed by SO_2 and O_3, both showing low values of positive contribution indicating the impact as mild but relevant as compared to other pollutants. In fact, ammonia was found to have high negative coefficients, perhaps reflecting an antagonistic influence on AQI at certain circumstances within the environment or some model characteristics. The whole picture painted by these coefficients indicates CO, NO, and NO_2 as significantly less important with respect to AQI, either of very small contribution or contextualized within their effects as far weaker features are concerned. One could assume it as giving even more quiet useful insights about describing direction of importance in understanding how each pollutant contributes itself with respect to air quality.

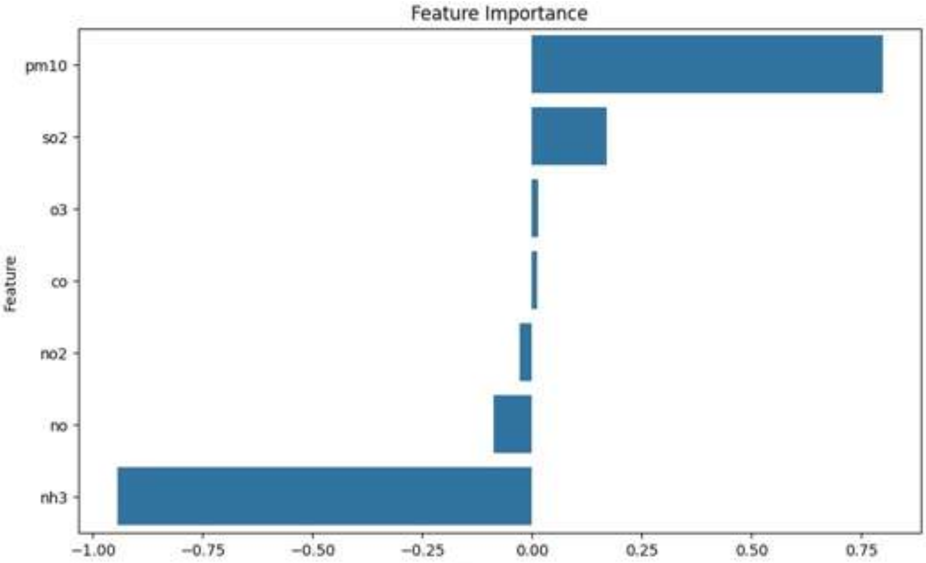


Figure 4: Feature Importance of Coefficient of Pollutants

4.5 Actual vs Predicted PM 2.5 Levels

The scatter plot entitled Actual versus Predicted PM2.5 Levels indicates how well the machine learning model predicts the PM2.5 concentrations based on the pollutant information. Each point in the plot marks one observation, with the x-axis indicating the actual PM2.5 value as per the dataset and the y-axis predicted PM2.5 values corresponding to observations. The points that cluster pretty tightly about the diagonal line indicate positive correlation coefficients between actual and predicted values suggesting that the model is fairly good at predicting PM2.5 concentration values. The very small deviations from the diagonal line depict prediction errors, but largely this model appears well captured and trustworthy enough to forecast PM2.5 levels.

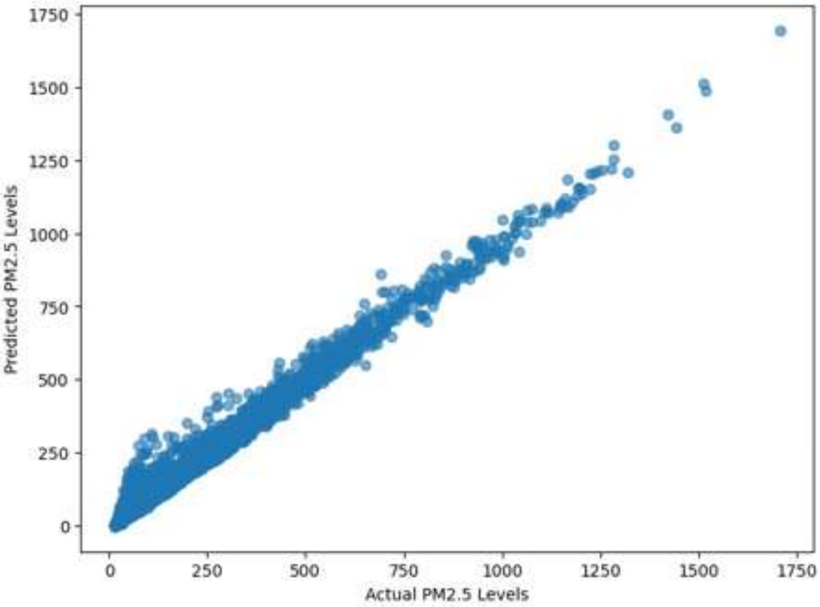


Figure 5: Actual vs Predicted PM 2.5 Levels

5. CONCLUSION

In this work, PM2.5 levels for Delhi will be forecasted using a machine learning technique in which different atmospheric pollutants are employed as the features. The use of Random Forest and XGBoost models, among others, facilitated the prediction of high accuracy in fine particulate matter concentrations. While they are usually based on descriptive statistics that define the variation and distribution of pollutant levels, correlation analysis demonstrates closer relationships for the most important parameters, especially PM2.5 with respect to PM10. Feature importance analysis concluded that two of the most important variables influencing the prediction process were PM10 and NH3. Furthermore, actual PA2.5 values matched closely to predicted ones. This finding will improve understanding of air quality dynamics while, at the same time, setting ground for real-time AQI forecasting and public health advisory in cities like Delhi.

6. REFERENCES

[1] Gupta, N. Srinivasa, et al. "Prediction of air quality index using machine learning techniques: a comparative analysis." Journal of Environmental and Public Health 2023.1 (2023): 4916267.

[2] Analitis, Antonis, et al. "Prediction of PM2. 5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods: A case study in London." Atmospheric Environment 240 (2020): 117757.

[3] Bhattacharya, Samayan, and Sk Shahnawaz. "Using machine learning to predict air quality index in new delhi." arXiv preprint arXiv:2112.05753 (2021).

[4] Natarajan, Suresh Kumar, et al. "Optimized machine learning model for air quality index prediction in major cities in India." Scientific Reports 14.1 (2024): 6795.

[5] Kumar, K., and B. P. Pande. "Air pollution prediction with machine learning: a case study of Indian cities." International Journal of Environmental Science and Technology 20.5 (2023): 5333-5348.

[6] Pradhan, Sushree Subhaprada, and Sibarama Panigrahi. "Studies on machine learning techniques for multivariate forecasting of Delhi air quality index." International Conference on Advances in Data-driven Computing and Intelligent Systems. Singapore: Springer Nature Singapore, 2022.

[7] Ameer, Saba, et al. "Comparative analysis of machine learning techniques for predicting air quality in smart cities." IEEE access 7 (2019): 128325-128338.

[8] Singh, Jayant Kumar, and Amit Kumar Goel. "Prediction of air pollution by using machine learning algorithm." 2021 7th International conference on advanced computing and communication Systems (ICACCS). Vol. 1. IEEE, 2021.

[9] Pal, Sudhanshu, Debanshi Pramanik, and Ena Jain. "Effectiveness of machine learning algorithms in forecasting AQI." 2021 International Conference on Technological Advancements and Innovations (ICTAI). IEEE, 2021.

[10] Chakravarty, Anwesha. "An Exploratory Analysis of Delhi Air Quality Using Statistics and Machine Learning Models." 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). IEEE, 2022.