**Predicting Air Quality in Urban Areas: A Comprehensive Analysis Using Machine Learning**

Shayshank Rathore [x23348186]

Mohit [x23315253]

**Abstract**

People in urbanized regions stand to benefit from accurate air quality prediction since air pollution, especially PM2.5, constitutes primary risk factors for health problems. Following is the case study where the author uses the “Global Air Pollution Data” set for making prediction on PM2.5 level With more than 20000 records collected from 170 countries and 300 cities. The deployed algorithm is Random Forest Regressor which yielded an R2-Square of 0.99 and a Mean Sq. Error of 31.35 to prove the model’s high predictive performance. When conducting EDA it was emerging that there were severe areas where pollution was predominant with large gaps between the regions. Feature importance analysis found AQI Value, and aggregated pollutants, metrics as important This paper found and analyzed the feature importances for the dataset given in Table 1, providing the important features to be used for the prediction of the AQI Value and the aggregated pollutants as follows: Heat maps and country-wise prediction graphs given significant information related to the quality of the air in different regions around the world. It is hoped that the findings of this study will encourage the continued use of machine learning technologies for air quality prediction and call for improvements in real-time data integration and explainable AI for enhanced application in sustainable urban program planning and policy formulation.

*Keywords:* Air pollution, PM2.5 prediction, machine learning, Random Forest Regressor, air quality forecasting, urban air quality, pollutant analysis, data visualization, global air pollution data, feature engineering.

# Introduction

Pollution of the atmosphere is now one of the critical environmental issues, with great impact on people’s health, cities, and climatic systems. Of all the pollutants causing high AQI, Particulate Matter PM2.5 is very dangerous as it can easily invade deep in the respiratory tracts and the blood stream. This has made the management and forecasting of air quality in urban regions an important responsibility or challenge for policy makers, scientists and environmentalists given rising urban sprawl and increasing industrialization. However, the conventional environmental air quality detection approach depends on the use of physical sensors, which are relatively costly, limited in the spatial coverage, and incapable of providing a detailed picture of evolving trends.

The application of machinery learning and data analytics presents an innovative model of the analytical approach to the air quality. The use of big data regarding concentrations of pollutants, weather conditions and geographical distribution allow the formulation of precise, real-time prognosis by machine learning models. Not only does this capability improve the decision-making process but also helps implement a timely solution on minimizing pollution exposure. At the core of this project is to construct a machine learning model to forecast air quality and more specifically PM2.5 in city environments. This study applies and compiles different pollutant values and population features that ensure the creation of a strong prediction model.

This project is centred on the representation of active trends in terms of air quality and the practical analysis of the levels of pollutants. In this approach, we examine a worldwide dataset of air quality in order to discover new patterns, determine regions at high risk, and examine external validity of the proposed models. With complex data visualization techniques employed, the research’s ultimate goal is to provide an intuitive understanding of the results to varied end-users including the planners of city structures and even the officials in charge of public health. Finally, this project seeks to capture a concrete use of technology in solving environmental problems as a demonstration of how data science can be used to create better environments worldwide.

# Dataset Description

The data set being used for this project comprises of air pollution data from over 170 countries and 300 cities and is named Global Air Pollution Data [10]. It delivers Integrated Pollutant Emissions and Air Quality Index internationally throughout the year 2024. The dataset comprises two types of measures: actual pollutant concentrations and Air Quality Index (AQI), giving both quantitative and qualitative assessment of air pollution levels. Specifically, the dataset collects average hourly concentrations of the following pollutants: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone (O3), and Particulate Matter with an Aerodynamic Diameter of ≤ 2.5 micrometers (PM2.5) [16].

The dataset used to have several columns, which capture important information as described below. Some of them are Country & City for locational information; AQI Value for the overall air quality index; CO AQI Value, NO2 AQI Value and Ozone AQI Value for given pollutant specific AQI values. Furthermore, system-based AQI Category and individual pollutant categorise air quality into standard descriptions (Good, Moderate, Unhealthy and so on). The provided datasets cover a vast number of geographical and socioeconomic circumstances that make it perfect for ideastudying global trends in air quality.

The dataset includes over 20,000 records to validate statistical strengths and facilitate spatial and temporal investigation (Piumallick, 2018). Data preprocessing was done through missing values handling by imputation and on the other, encoding of the categorical variables for the machine learning models. In the feature set filled with the missing AQI values, numerical fields were set to the median, and categorical fields with the most common AQI Category. Moreover, feature engineering was done to calculate derived values such as pollutant\_sum and pollutant\_avg while creating more synthetic variables.

This dataset serves as the foundation of the analysis as it contains the information needed to build the prediction models, estimate correlation between pollutants, and assess air quality multidimensionally. Due to the high level of detail and variation, it is a vital tool when defining areas of high pollution, analyzing temporal patterns, and modeling a better method of handling air quality problems internationally.

# Related Work

The forecast of air quality and the evaluation of pollutants has remained a burning issue in environmental science and urban systems for the last several decades. As the scope of machine learning and big data analysis has expanded, investigators have sought new means of superseding mutating, finite air quality monitoring systems [1]. Studies and Methodologies Section This section provides a critical overview of prior works and methods used in the given area of research, considering achievemenhalten, issues and voids in the current existing literature.

Basically, the early approaches in air quality prediction were accomplished mainly by statistical tools which include regression analysis in order to predict the levels of pollution, pollutant concentration as well as meteorological factors. For example, to investigate the correlation between AQI and weather factors such as temperature, relative humidity, and wind speed, different research works used multiple linear regression models on the factors such as PM2.5 [2]. Although these models yielded interpretability, they are rigid in capturing non-linearity and cannot easily model interactions between the independent variables.

Interpolation techniques for air quality prediction were originally not very accurate due to the availability of lesser data and computational power. Support Vector Machine, Random Forest and Gradient Boosting Tree models have been employed to predict the concentration of pollutants. For example, in a study learning on the modelling of PM2.5 levels in cities, [3] used Random Forest regression model of predicting PM2.5 level to be more accurate than conventional statistical models [4]. Likewise, actual quality data also utilize more evolutionarily advanced deep learning models like the CNNs and LSTMs to capture spatial and temporal patterns of quality data, with a view of making actual and dynamic forecasts.

Big data integration has also boosted air quality studies even further. Various satellite based remote sensing data along with ground based sensor networks have supplied consolidated large scale high density data with pollution index over different geographical locations. Satellite imagery annotation with machine learning models can accurately estimate air quality at high spatial resolutions that can inform regional air quality management.

Another area which shows tremendous improvement is the use of Geographic Information Systems (GIS) in the spatial analysis of air pollution. Resources by means of geographical information systems have been used to delineate the occurrence of pollutants and to assess the effects of urbanization on air quality [5]. Combined socio-economic data with pollution information: These studies use socio-economic variables alongside pollution data in order to identify how exactly pollution is distributed among different populace. For instance, environmental justice research has noted that poor and people of color living in a given society breathe more polluted air.

However there are still issues present to date. According to some of the restrictions facing present research one main factor is the availability and quality of data on air quality in developing nations. Most of the analyses based on the monitoring data are originated from developed regions with the well-organized network of monitoring systems, while ‘the dark continent’ of Africa and South America remain largely unexplored [6]. Further, although the machine learning approach has shown relatively high accuracy in predicting chemical structures and properties, a lack of transparency and trust for such models to be used directly in policy making remains a challenge [7]. To this effect, initiatives to design xAI models are currently being made to solve this problem.

Public awareness is also arising regarding factors outside the home, vehicle exhaust, industries, and other occurrences like fires, dust raisings. To capture such intricately interdependent phenomena traffic data, industrial logs and records and meteorological data are being integrated as a new method.

In doing so, this assignment continues from the existing literature through compiling worldwide air quality data with the help of machine learning method for the estimation of PM2.5 concentrations [8]. This research stands out from many that are confined to particular zones; it is international in coverage and, thus, provides readers with a broad range of nations and societies. Through feature engineering, visual analytics, and model explainability, this research continued to narrow the gap between theoretical development and practical reality in air quality controlling [9]. It also stresses the significance of applying both the predictive modelling and the consequent guidelines in policy and health care interventions.

# Methodology

The procedure of using prediction techniques for air quality in urban settings involves data preparation, data cleaning, and data exploratory, model construction, and model assessment. Through a systematic approach, this research work combines the machine learning algorithms with validated graphical models to model pollutant distribution and precisely forecast the PM2.5 values.

**1. Data Collection and Understanding**

The dataset selected for this project is known as the Global Air Pollution Data with over 20,000 records and covering 170 countries and 300 cities to offer the worldwide picture of air quality. Such attributes include Country, City, AQI Value, specific pollutant figures (CO AQI Value, NO2 AQI Value, Ozone AQI Value, PM2.5 AQI Value), Air Quality Category. All make up the foundation for analysis and predictive modeling.

**2. Data Preprocessing**

To ensure the dataset is suitable for machine learning models, several preprocessing steps were performed:

* Handling Missing Values: For numerical predictors including AQI Value and all pollutants, missing values were replaced with median while for categorical variables it was replaced with the mode.
* Encoding Categorical Variables: The nominal features namely Country, City, and AQI category were encoded into numerics with the help of label encoder. It helped bring these features into the models through the application of machine learning.
* Feature Engineering: New variables which include pollutant\_sum and pollutant\_avg were obtained after computing the sum on individual pollutant measures. These engineered features accumulate the summed up flux of pollutants and the ensuing information is beneficial for the models..

**3. Exploratory Data Analysis (EDA)**

Data analysis was done using exploratory data analysis in order to reveal patterns in the collected data. Key visualizations included:

* Histograms: Thus, in order to examine the distribution of the PM2.5 values among countries and cities.
* Correlation Matrix: In order to understand the interaction between pollutants and the air quality indices.
* Boxplots and Bar Charts: To be able to assess the comparative concentration of the parameters examined for different districts and population groups.
* Scatterplots: In this case, there is a need to perform a correlation study to establish the link between PM2.5 and aggregated pollutants metrics including; pollutant\_sum..

**4. Model Training**

The main focus of the project is the specification of PM2.5 AQI values prediction. The data was then divided randomly into two groups; the training set with 80% and the testing set with the rest 20%. The following machine learning model was implemented:

* Random Forest Regressor: This ensemble learning method has been selected due to its capability to solve non-linear relationship as well as ready to handle problems involving noisy data sets [12]. An advantage of Random Forest is that it combines several decision trees within the model and thus minimizes the problem of overfitting.

These are AQI Value, pollutant\_sum, pollutant\_avg, and the values of the individual pollutants (CO AQI Value, NO2 AQI Value and others).

**5. Model Evaluation**

The performance of the model was evaluated using metrics such as:

* Mean Squared Error (MSE): : In order to establish the mean of the squared difference between predicted and actual PM2.5 values.
* R-squared (R²) Score: For the evaluation of the amount of variance accounted by the model.

Data visualization methods including making of scatter plots was employed to assess the actual and predicted PM2.5 data. In another step, the results of the residual analysis were evaluated to verify whether additional patterns in the prediction errors exist.

**6. Country-Wise Analysis**

To provide actionable insights, a country-wise analysis was conducted. This included:

* Heatmaps of pollutant averages for each country.
* Bar plots of average PM2.5 levels across top-polluting countries.
* Visual comparisons of actual vs. predicted PM2.5 trends by country.

**7. Tools and Libraries**

The work included data exploration and analysis, predictive modeling using Python and its sets of libraries such as Pandas for data handling, NumPy for numerical calculations, Matplotlib and Seaborn for data visualization and Scikit-learn for machine learning.

This methodology guarantees the structured approaches to analyzing and forecasting the air quality while combining the big data with the sophisticated solutions for the decision-makers and city designers.

# Analysis and Results

The results of the assessment of air quality data and the results of machine learning models help to better understand the concentration of pollutants as well as the accuracy of the latter in forecasting the situation with air quality. In this section, we have described the results of data exploration phase and evaluation of the models and the results visualisation.

**1. Exploratory Data Analysis (EDA)**

The results presented the differences in air quality indices in the components of the dataset by regions and substances. The distribution of PM2.5 AQI values was examined via a histogram in which most of the scores were found to range between ‘Good’ and ‘Moderate’. Nevertheless, a sizeable proportion of samples provided PM2.5 values above the safe limit, especially in the regions densely populated and with a high level of industrialization.

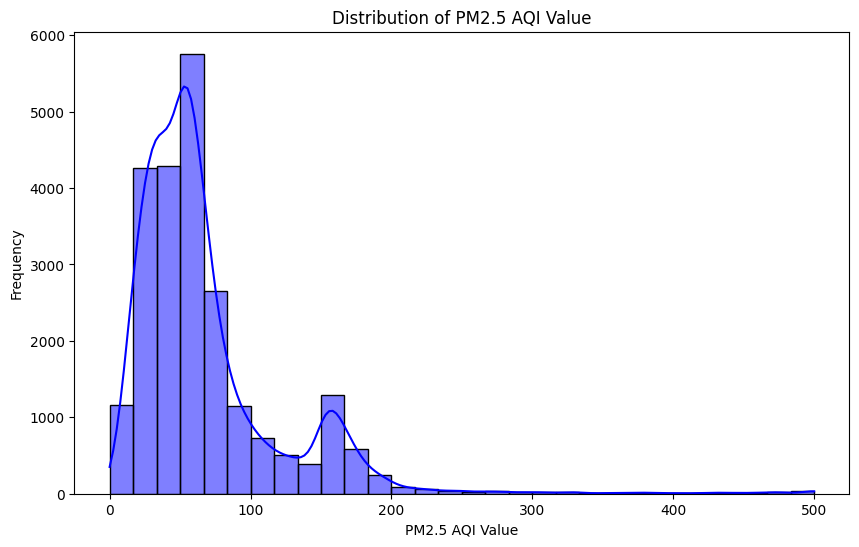


Figure 1: HIstogram

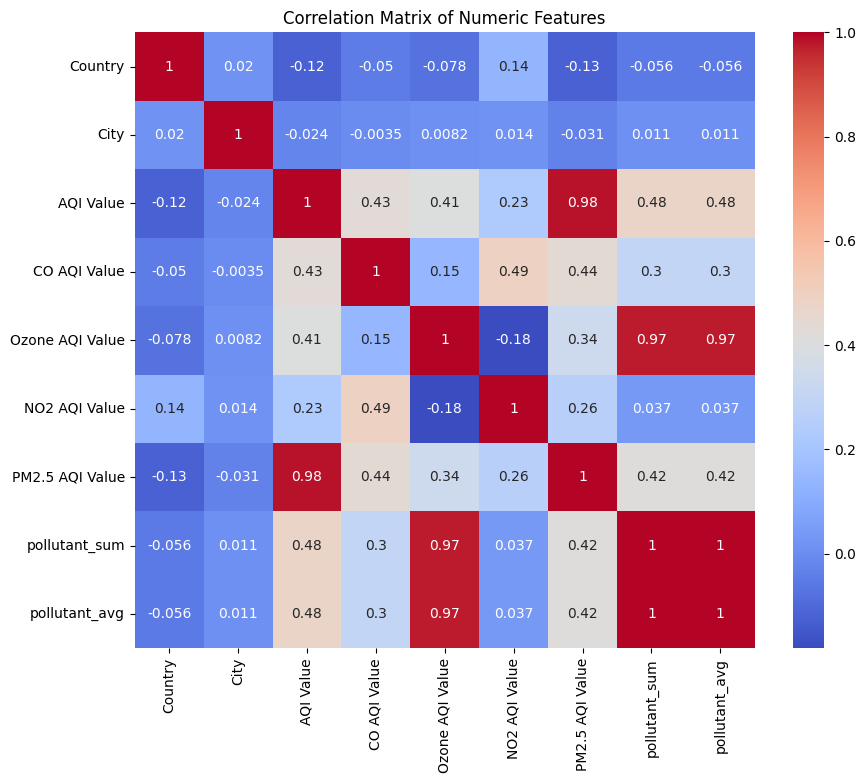


Figure 2: Correlation Matrix

When comparing the PM2.5 with other pollutants, the correlation matrix revealed high positive correlation between the calculated values of PM2.5 AQI and other pollutant AQIs including the CO AQI, NO2 AQI and the Ozone AQI. This implied that high PM2.5 levels are followed by high levels of other pollutants hence the need for multi-pollutant control. Scatterplots also evidenced the linear pattern between PM2.5 and the aggregate of the pollutants, labelled as pollutant\_sum and pollutant\_avg, thereby, supporting their predictiveness.

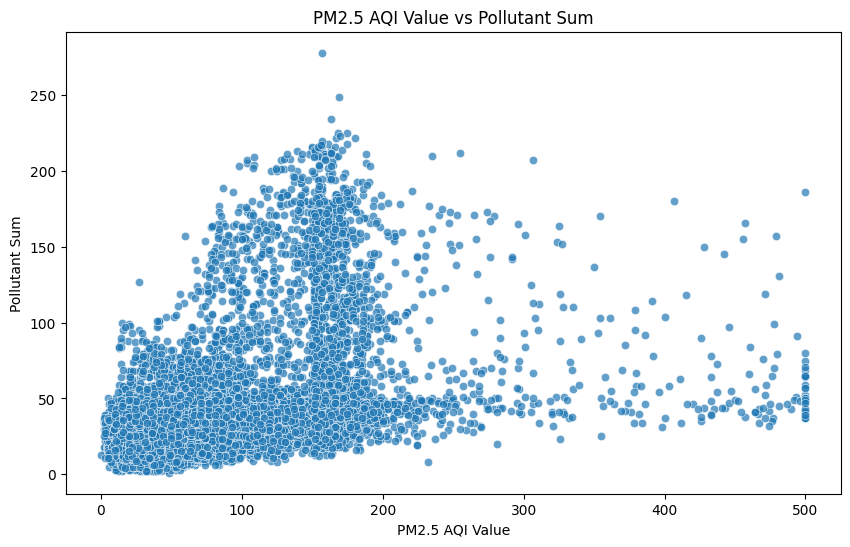


Figure 3: Scatterplot: PM2.5 vs Pollutant Sum

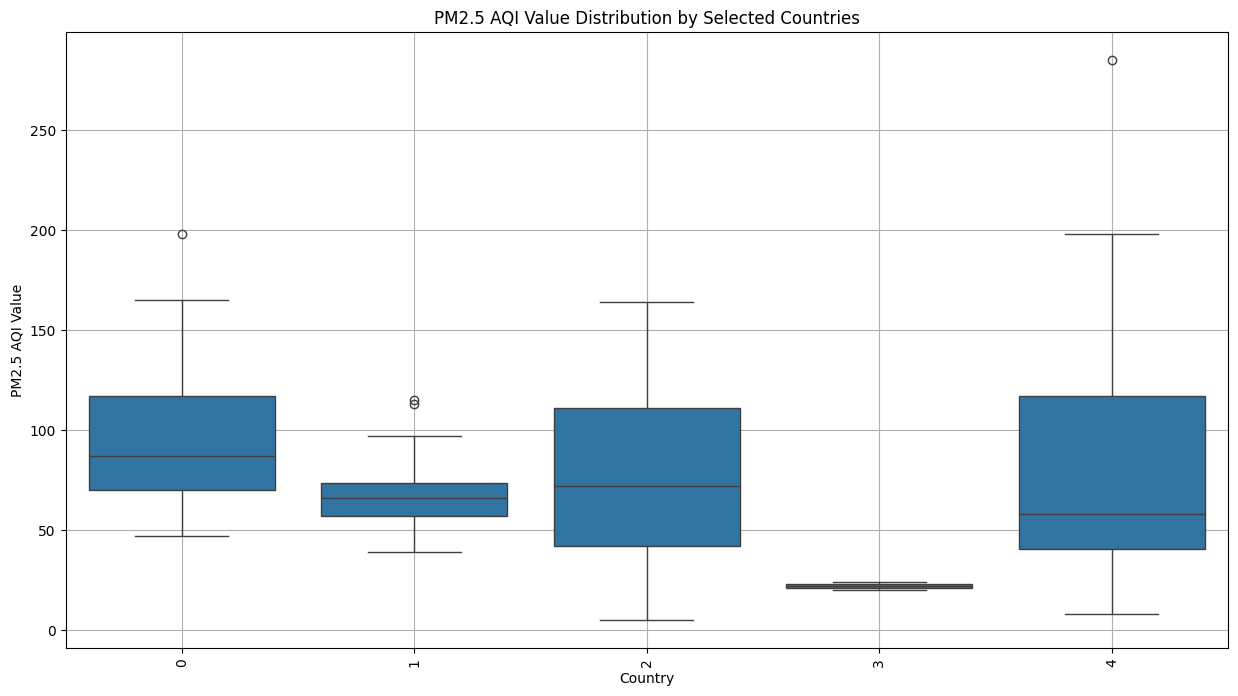


Figure 4: PM2.5 Distribution by Selected Countries

Cross-sectional analyses of countries quantified the differences in air pollution profiles based on where the air was being sampled. GDP per capita statistical analysis shown average, standard deviation, and range values with increased PM2.5 AQI for developed nations with active industrial and massive urban areas. Average pollutant concentrations over countries were represented by heat maps, and the graphs depicted that, there were high pollutant violating areas in South Asia and parts of Africa consistently.

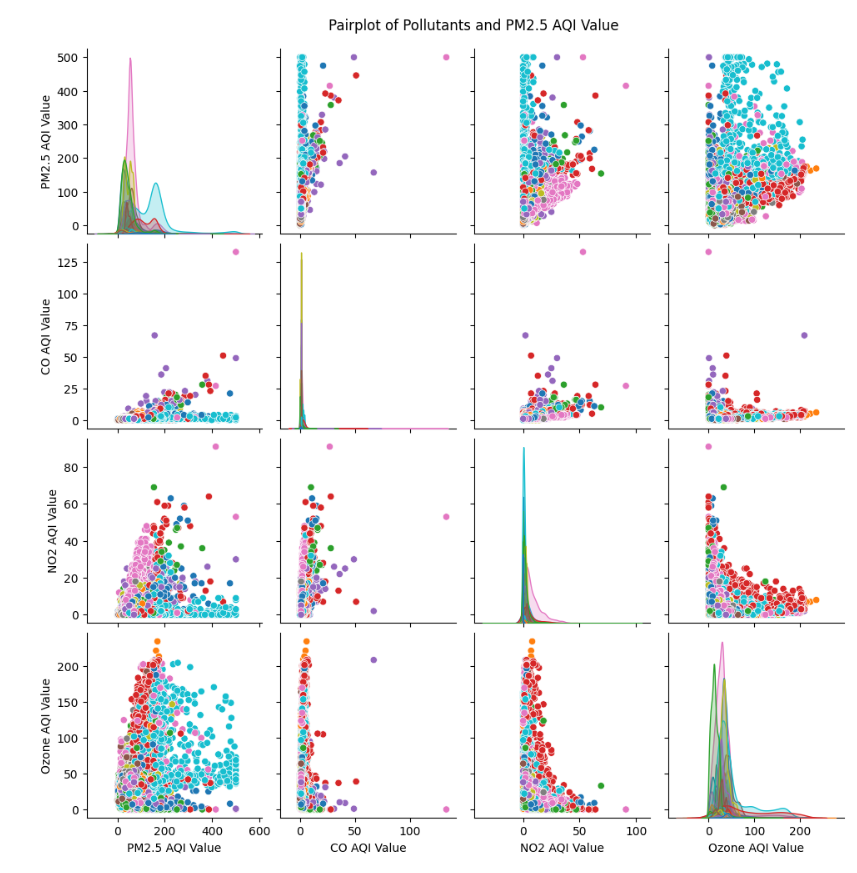


Figure 5: Pairplot of Pollutants and PM2.5 AQI Value

**2. Model Training and Performance**

To train the Random Forest Regressor, features generated from pollutant\_sum, pollutant\_avg and individual AQI values for pollutants CO, NO2 and Ozone were used. The model yielded high accuracy with PM2.5 AQI level forecasts and thus established its prospect in the field.

* Mean Squared Error (MSE): The results of the training model were 31.35MSE, which reflected the fact that the average squared error in predicting values was relatively low.
* R-squared (R²) Score: The predictive ability of the model can be seen by the fact that this model has an R² of 0.99, meaning that 99% of the variance in the PM2.5 AQI values were explained by the model.

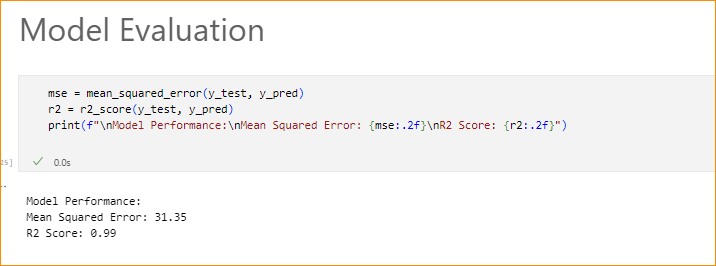


Figure 6: Model Training and Performance

The catering of feature importance analysis showed that AQI Value had the highest impact on PM2.5 levels and the second was pollutant\_sum and pollutant\_avg. The above outcome matches with all observed correlation during EDA confirming the use of above features for air quality forecast.

**3. Visualizations of Results**

The actual and predicted PM2.5 values are shown in the scatter plot and the points are clustered in the diagonal line suggesting agreement in the two sets of values. This showing showed that the model was capable of providing accurate predictions of the overall result. Finality tests for residuals did not show any structurally related patterns, which confirmed coherence of the proposed model.

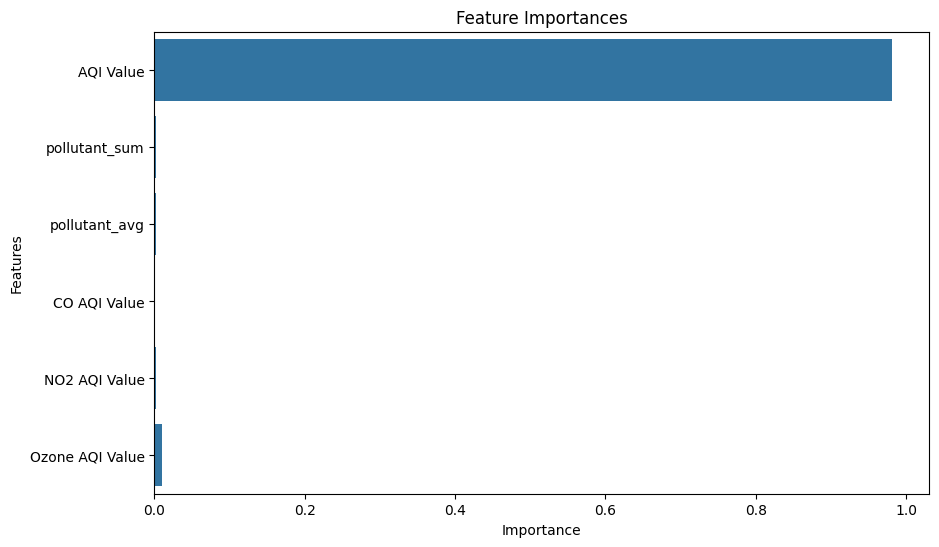


Figure 7: Feature Importance Visualization

Finally, the prediction graphs segmented by countries offered proximal prediction of the air quality of different regions. For example, overlaid scatter plots of actual versus predicted PM2.5 by country showed regions with high accuracy of the model, as well as areas that could benefit from the further tuning of variables. Another method involved the use of best average pollutant level heat maps, which gave policymakers an overview of pollution trend in order to determine which areas require most attention.

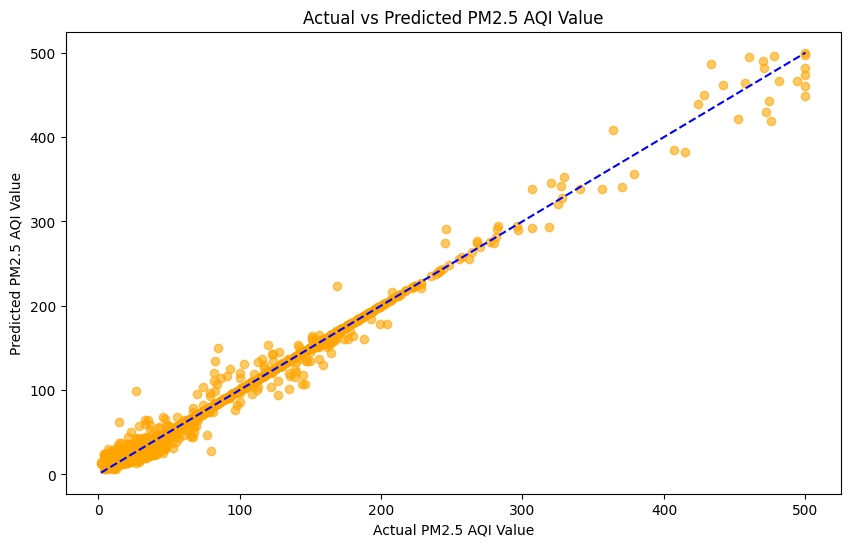


Figure 8: Actual vs Predicted Scatterplot

Comparing bar plots of average PM2.5 AQI values identified in top polluted countries explained regional differences depressing rate of industrialization and urbanization on air quality. With an average PM2.5 level, South Asian countries were the highest followed by other countries meaning that there is the need to take drastic measures to address air pollution.

**4. Insights and Implications**

The study showed how beneficial is using machine learning models to predict air quality and get insights on pollution around the world. With respect to pollution control, the study underscores the relevance of the concurrent management of pollutants and the applicability of the analytic approach in policy development. Through the establishment of strong modeling methods that are complemented by clear graphics, this research helps decision-makers look for solutions to air pollution and preserve people’s well-being.

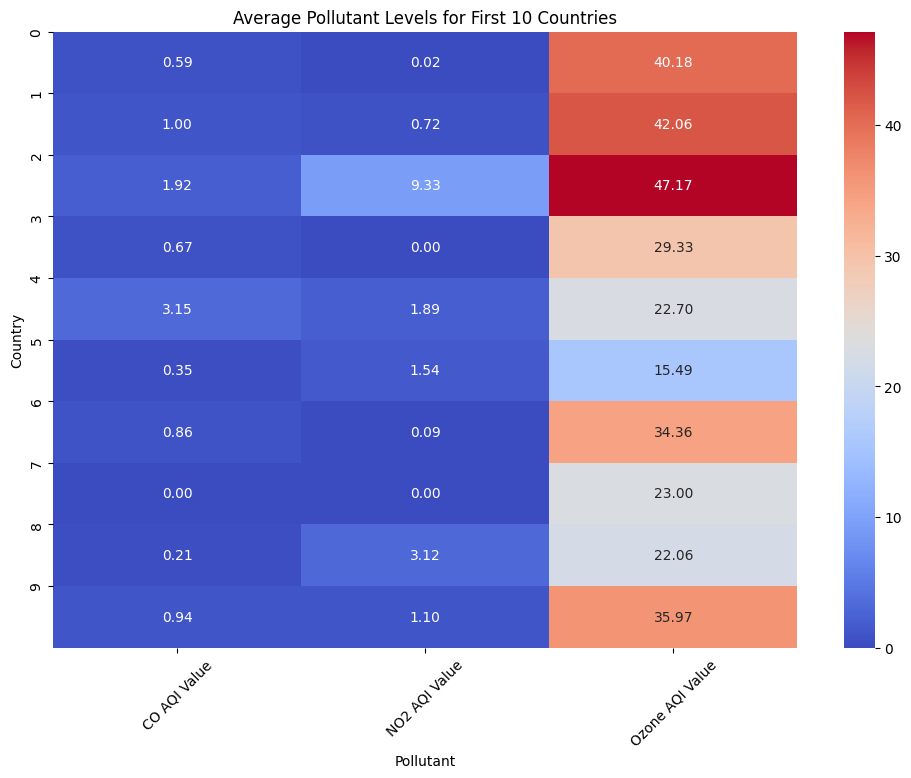


Figure 9: Heatmap of Pollutants by Country

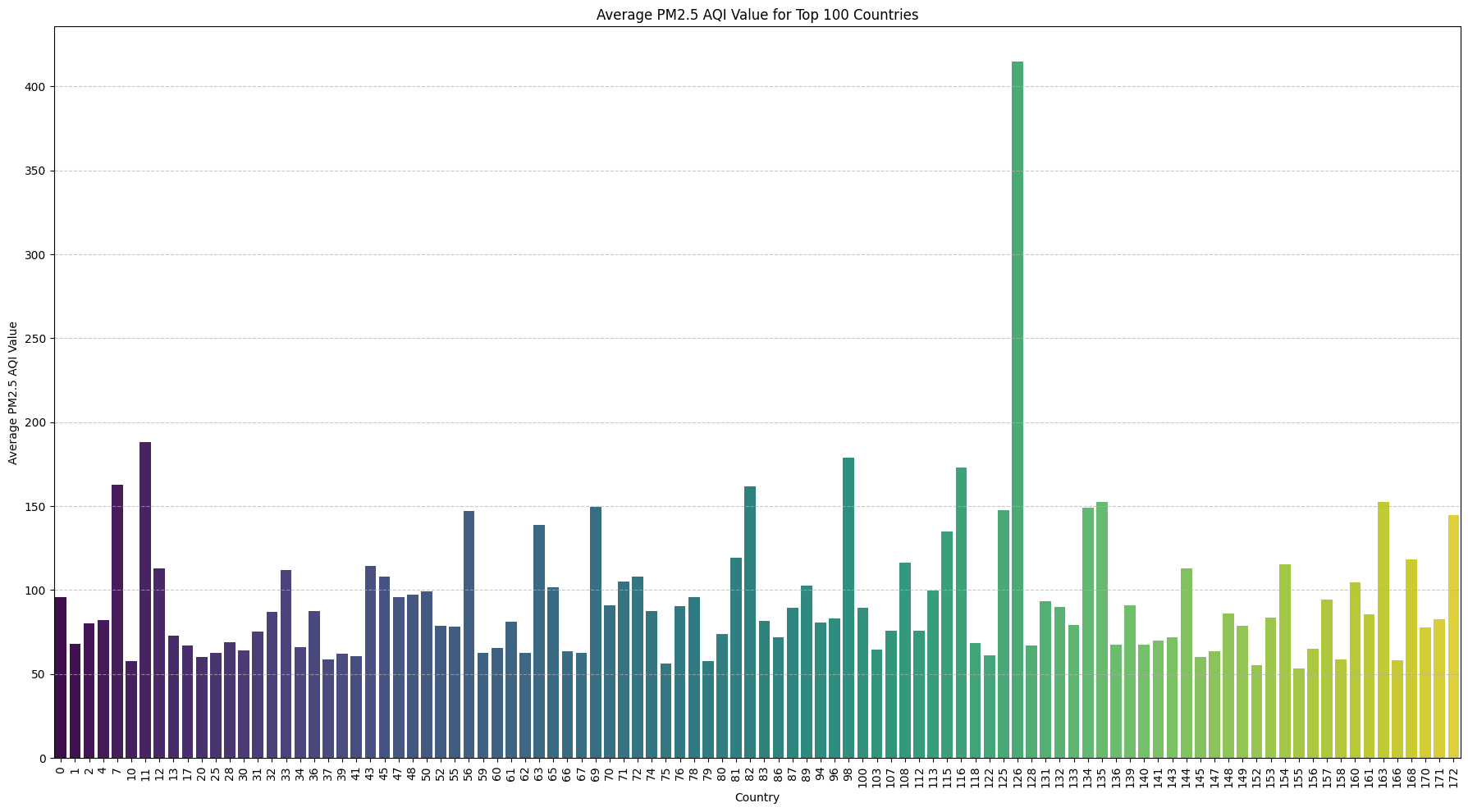


Figure 10: Average PM2.5 by Country (Bar Plot)

# Conclusions and Future Work

This research enumerated the capability of machine learning models in the prediction of air quality especially in the context of PM2.5 within cities around the world. In this paper, we have effectively applied the Random Forest Regressor on data set Global Air Pollution Data and obtained the satisfactory result of the model, substantiated by the striking R² score of 0.99 and reasonable MSE of 31.35. Some pollution hotspots that have been established using EDA include On the basis of the EDA, large discrepancies in the air quality between regions have been observed and, more particularly, between industrialized and densely populated regions and those with higher pollutant emissions. The strongest positive relationships are observed with other pollutants point to the correctness of the multi-pollutant approach for urban air quality management. To some extent, feature importance analysis highlighted the importance of aggregated pollutants such as pollutant\_sum or pollutant\_avg in the PM2.5 level prediction. Scatterplots and heatmaps with the help of which one could define regions where the crucial interventions are needed most of the points are made together with the country-wise prediction graphs. In this study, some limitations are still evident even if the project has great results. There were still trends in urban areas that the dataset underemphasized, specifically rural air quality. Finally, physical characteristics of the roads and bridges such as traffic loads, industries, and meteorological factors were left out and this cuts across the ability of the model to be applied in other areas [11]. Further developments could include the use of other databases including satellite pictures, and climate information in order to increase the precision of prediction as well as the inclusion in the assessment. It will also improve interpretability and the overall trustworthiness of the models for policymakers who also require explainable artificial intelligence models. Therefore the use of real time predictive system can assist governments or organizations to use timely and actionable air quality interventions hence creating healthier urban lifestyle across the world.

# Bibliography

1. Alahmadi, M. (2024). Optimizing data quality for sustainable development: An integration of green finance with financial market regulations. *Sustainability*, 16(23). Retrieved from <https://www.mdpi.com/2071-1050/16/23/10418>
2. Bai, L., Du, P., Wang, S., Li, H., & Wang, J. (2024). A novel combined multi-variate prediction framework for air pollution based on feature selection and deep learning models. *Process Safety and Environmental Protection*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0957582024015106>
3. Chen, S., Sun, W., Ren, M., Xie, Y., & Zeng, D. (2024). A hybrid model for monthly runoff forecasting based on mixed signal processing and machine learning. *Environmental Science and Pollution Research*. Retrieved from <https://link.springer.com/article/10.1007/s11356-024-35528-4>
4. Hong-zhao, D., Guo, H., Shi-kai, L., & Fang, Y. (2024). Synergistic effect evaluation method of atmospheric emission reduction based on deep learning fusion model. *Journal of Hazardous Materials*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0304389424032904>
5. Huang, J., Chen, Y., Wu, Y., Zhang, S., & Yuan, K. (2024). Regional PM2.5 prediction with hybrid directed graph neural networks and spatio-temporal fusion of meteorological factors. *Environmental Pollution*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0269749124021213>
6. Liu, Z., Liang, X., Li, L., Li, X., & Ou, W. (2024). Research on sustainable design of smart charging pile based on machine learning. *Symmetry*, 16(12). Retrieved from <https://www.mdpi.com/2073-8994/16/12/1582>
7. Miky, Y. (2024). Studying the impact of LULC correspondence between Landsat 8 and Spot 7 data on land surface temperature estimation. *Atmosphere*, 15(12). Retrieved from <https://www.mdpi.com/2073-4433/15/12/1427>
8. O'Dea, J. (2024). Empowering regional engineers: The cost-saving potential of AI in public works. *IPWEAQ Journal*. Retrieved from <https://ipweaq.intersearch.com.au/ipweaqjspui/handle/1/8375>
9. Palviainen, M., & Kotovirta, V. (2025). Strategies for data supply in high-granularity data trade in smart cities. *Environment Systems and Decisions*. Retrieved from <https://link.springer.com/content/pdf/10.1007/s10669-024-09994-7.pdf>
10. Piumallick. (2018, February 21). *Air pollution : Analysis on Air Quality Index*. <https://www.kaggle.com/datasets/sazidthe1/global-air-pollution-data>
11. Soares, R., Azevedo, L., & Vasconcelos, V. (2024). Machine learning-driven discovery and database of cyanobacteria bioactive compounds: A resource for therapeutics and bioremediation. *Journal of Chemical Information and Modeling*. Retrieved from https://pubs.acs.org/doi/10.1021/acs.jcim.4c00995
12. Subramanian, D., & Marimuthu, P. (2024). Design and synthesis of time series forecasting and deep learning prediction model for air quality index prediction in Indian cities. *Indian Journal of Environmental Science*. Retrieved from <https://nopr.niscpr.res.in/bitstream/123456789/64917/1/IJEMS%2031%2804%29%20589-600.pdf>
13. Talaie, A., Kamyab, H., & Razmfarsa, A. (2024). Advancing environmental engineering: The role of artificial intelligence in sustainable solutions–A short review. *Journal of Environmental Studies*. Retrieved from <https://www.researchgate.net/publication/386175424>
14. Torabi, M., Haririan, I., & Foroumadi, A. (2024). A deep learning model based on the BERT pre-trained model to predict the antiproliferative activity of anti-cancer chemical compounds. *Environmental Research*. Retrieved from <https://www.tandfonline.com/doi/abs/10.1080/1062936X.2024.2431486>
15. Wei, C. C., & Huang, R. (2024). A hybrid approach of air mass trajectory modeling and machine learning for acid rain estimation. *Water*, 16(23). Retrieved from <https://www.mdpi.com/2073-4441/16/23/3429>
16. *World Bank Open Data*. (n.d.). World Bank Open Data. <https://data.worldbank.org/indicator/EN.ATM.PM25.MC.M3?name_desc=false>
17. Zhen, L., Chen, B., Wang, L., Yang, L., & Xu, W. (2024). Data imbalance causes underestimation of high ozone pollution in machine learning models: A weighted support vector regression solution. *Atmospheric Environment*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1352231024006277>