**PREDICTIVE ANALYSIS OF AIR POLLUTION IN URBAN AREAS USING REGRESSION MODELS**

Hein Thuya Win, Shuvanga Bikram Hamal, Phyo Thwe Thwe Kyaw

Department of ICT, Rangsit University

*\*Corresponding author:* [*heinthuya.w65@rsu.ac.th*](mailto:heinthuya.w65@rsu.ac.th)

**Abstract**

Air quality is a critical factor affecting the health and well-being of urban populations, making the accurate prediction of Air Quality Index (AQI) values essential for urban planning and public health interventions. This research explores the application of three regression models—Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost)—to predict AQI values based on a comprehensive dataset of major air pollutants, including PM2.5, PM10, NO2, SO2, CO, and O3, in urban areas of India from 2015 to 2020. The models were trained and evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). Among the models, XGBoost achieved the highest R² (0.9436), indicating superior predictive accuracy, followed by Gradient Boosting (0.9388) and Random Forest (0.9371). XGBoost also demonstrated the lowest MAE (13.91) and RMSE (21.31), reflecting its ability to effectively handle data variability and deliver more accurate predictions. These findings underscore the potential of ensemble methods in improving AQI prediction, providing valuable insights for urban air quality management and public health strategies. This study contributes to the growing field of environmental data mining and highlights the role of machine -learning models in advancing air quality prediction and monitoring.

***Keywords:*** *Air Quality Index (AQI), Random Forest, Gradient Boosting, XGBoost, Air Quality Monitoring*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **INTRODUCTION**

Air quality is an essential factor influencing the health, safety, and overall well-being of urban populations. In rapidly growing urban areas, air pollution has become a major public health concern, contributing to respiratory diseases, cardiovascular issues, and premature mortality. The Air Quality Index (AQI) serves as a vital tool for measuring and communicating the levels of pollutants in the air, providing a clear indication of the air quality at any given time. However, predicting AQI values with high accuracy is a complex challenge due to the dynamic nature of air pollution, which is influenced by a variety of factors such as weather conditions, industrial activity, traffic patterns, and geographical location.

Traditional methods of predicting AQI values often rely on direct measurements or simple statistical models, but these approaches can fall short in capturing the complex interactions among pollutants and environmental factors. With the rise of machine learning and data-driven approaches, more sophisticated methods, particularly regression models, have gained attention for their ability to predict AQI values based on large-scale, real-time environmental data. Regression models, such as Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), are particularly promising due to their ability to handle non-linear relationships, manage high-dimensional data, and provide more accurate forecasts than traditional models.

This research aims to explore the effectiveness of these regression models in predicting AQI values in urban areas, using a comprehensive dataset that includes major air pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3) from urban areas in India between 2015 and 2020. Given the complex and multifaceted nature of air pollution, understanding how well these models can predict AQI values could provide valuable insights into improving air quality forecasting, which is crucial for both public health and urban planning.

The primary research question guiding this study is: *How effectively can regression models predict AQI values based on major air pollutants, and what role can these predictions play in managing urban air quality?* In this context, the research investigates several key aspects:

1. **Model Performance:** How do different regression models—Random Forest, Gradient Boosting, and XGBoost—perform in terms of predictive accuracy and error metrics when applied to AQI prediction?
2. **Role of Pollutants:** How do various pollutants, individually and in combination, influence AQI predictions, and which pollutants have the greatest impact on AQI values?
3. **Practical Implications:** How can accurate AQI predictions, generated by these machine learning models, be used to inform urban air quality management practices, policy decisions, and public health interventions?

By addressing these questions, this research aims to contribute to the growing body of literature on environmental data mining and machine learning, offering insights into the potential of predictive modeling for improving air quality monitoring and management in urban areas. The results of this study could not only aid in the development of more reliable air quality forecasting systems but also provide a foundation for future research on the integration of AI-based tools in urban environmental policy-making.

This paper is organized as follows: The **Introduction** introduces the research topic and objectives. The **Literature Review** explores previous studies on AQI prediction using machine learning models. The **Methodology** outlines the dataset, models, and evaluation metrics used. The **Results and Discussion** compares the performance of Random Forest, Gradient Boosting, and XGBoost, analyzing their effectiveness in predicting AQI. Finally, the **Conclusion** summarizes the findings and suggests directions for future research.

1. **LITERATURE REVIEW**

The study of air quality and its impact on urban populations has gained significant attention in recent decades, particularly as urbanization accelerates globally. Urban air pollution, often caused by a combination of industrial emissions, vehicular traffic, and natural sources, has become a critical public health issue. The prediction and monitoring of air quality, particularly the Air Quality Index (AQI), is essential to mitigate health risks and inform environmental policies. This literature review explores the use of various machine learning models, especially regression techniques, in predicting AQI values based on air pollution data. The review also highlights the role of key pollutants, their influence on AQI predictions, and the potential applications of these predictions in managing urban air quality.

Over the last two decades, machine learning has become a key tool in environmental science, especially for air quality prediction. While traditional methods like linear regression and time series forecasting have been widely used, they often struggle to capture the complex, non-linear relationships between pollutants and AQI. In contrast, machine learning models, including Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost), excel at handling large datasets, identifying key features, and modeling complex, non-linear relationships. Random Forest improves accuracy by using multiple decision trees and averaging their predictions, while Gradient Boosting combines weak learners to iteratively enhance predictions. XGBoost, an optimized version of Gradient Boosting, stands out for its superior performance, efficiency, and ability to handle large-scale data, making it a popular choice for air quality forecasting.

The prediction of AQI is influenced by the types and concentrations of pollutants, particularly PM2.5, which is a major contributor to high AQI values and respiratory diseases in urban areas. PM2.5 can penetrate deep into the lungs, aggravating health conditions. Other pollutants like NO2, SO2, CO, and O3 also play a significant role, especially in areas with heavy vehicular traffic or industrial activity. For example, NO2 is a precursor to ozone formation, and O3 levels can lead to smog formation, both of which significantly impact AQI. Machine learning models, particularly ensemble methods like Random Forest and Gradient Boosting, are well-suited for modeling these complex, non-linear relationships. They can capture interactions between pollutants, such as the combined effects of high PM2.5 and O3, which may not be detectable by simpler models.

Accurate AQI prediction has important applications for urban management, including issuing pollution alerts, guiding public health recommendations, and optimizing resource allocation (e.g., air quality monitoring stations). In the long term, AQI forecasting can support urban planning by revealing how factors like traffic and industrial emissions affect air quality. This information can help inform policies aimed at creating healthier, more sustainable cities. Moreover, by offering timely predictions, these models could support decisions regarding traffic control and industrial emission reductions during high pollution events, improving real-time urban air quality management.

Although significant research has been done on AQI prediction using machine learning, several gaps remain. First, more region-specific studies are needed, particularly in developing nations like India, where air pollution varies greatly across different regions. Factors such as local climate conditions, topography, and seasonal variations can all influence pollutant concentrations and their impact on AQI. Second, while many studies assess individual machine learning models, few compare multiple regression techniques, such as Random Forest, Gradient Boosting, and XGBoost, using the same dataset and performance metrics. Such comparative studies would provide clearer insights into the strengths and limitations of each model in different contexts. Lastly, while AQI prediction has been applied in public health and urban planning, its integration into real-time decision-making systems—such as dynamically adjusting traffic flow or deploying mobile air quality sensors—is still underexplored.

In conclusion, regression models like Random Forest, Gradient Boosting, and XGBoost show promise for improving urban air quality management by accurately predicting AQI values. These models’ ability to handle complex relationships between pollutants and environmental factors makes them valuable tools for forecasting air quality and informing policy decisions. However, further research is necessary to address regional air pollution variations, better understand pollutant interactions, and explore the real-time application of AQI predictions in urban planning and public health management. This research aims to bridge these gaps by providing a balanced focus on both accuracy and interpretability through a comparative analysis of different models, as well as investigating the potential for real-time AQI prediction systems in urban environments.

1. **METHODOLOGY**

### ***3.1 Data Collection***

The dataset used in this study is publicly available on Kaggle and contains hourly and daily air quality data from monitoring stations across 25 cities in India, offering a broad representation of urban air quality. It includes key pollutants such as CO, O₃, NO₂, PM10, and PM2.5, along with corresponding Air Quality Index (AQI) values. With over 29,000 records, it provides valuable insights into pollution trends across different regions in India.

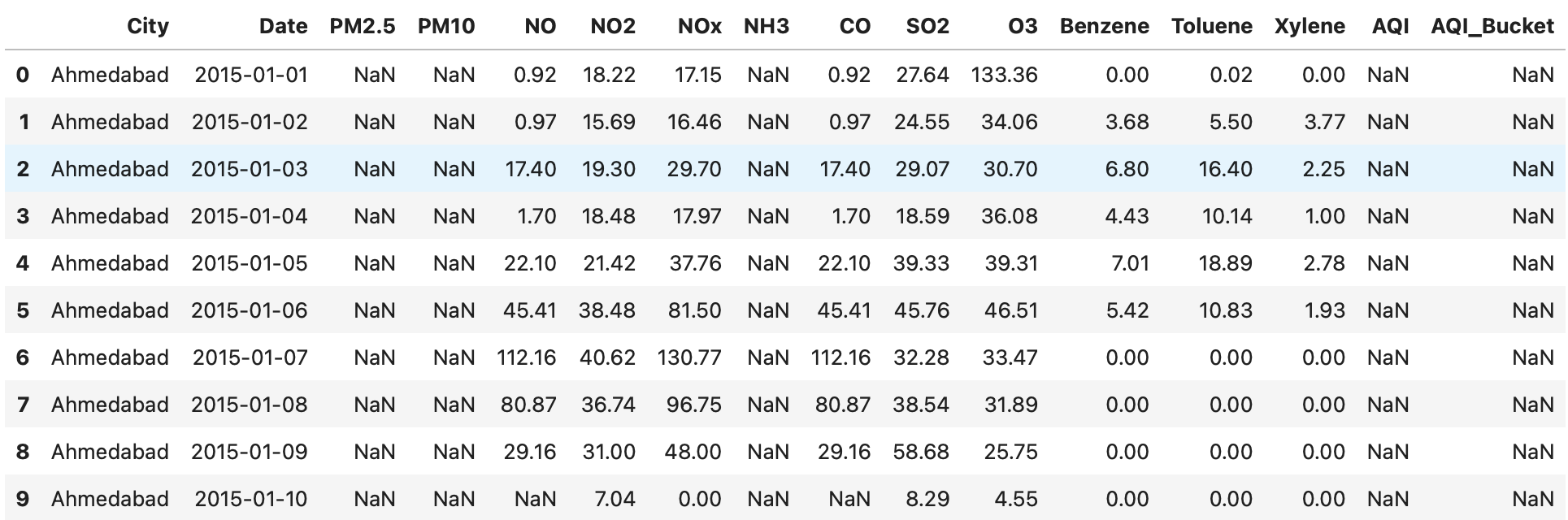


Figure 3.1: Dataset

### ***3.2 Data Preprocessing***

Data preprocessing is a crucial process in transforming the raw, inconsistent data into a clean, structured format suitable to achieve a more accurate and reliable AQI prediction model. The preprocessing steps were as below:

* **Handling Missing Data**: Missing values were handled by imputing missing values with the median of each respective column. This approach ensures that the imputation process does not introduce biases into the dataset.
* **Data Transformation**: No log transformation was applied, but the dataset was normalized where necessary. Features were standardized to ensure that they are on a comparable scale, which is particularly important for gradient-based models like Gradient Boosting and XGBoost.
* **Feature Engineering**: The features used for prediction include pollutant concentrations. The Date and AQI\_Bucket columns were excluded since they were not relevant to the regression task.



Figure 3.2: Preprocessed Data

### ***3.3 Model Selection***

For predicting the Air Quality Index (AQI) based on air pollution data, three powerful machine learning algorithms were selected: **Random Forest Regressor**, **Gradient Boosting Regressor**, and **XGBoost Regressor**. These models were chosen due to their ability to handle complex, non-linear relationships between various pollutants and AQI values, which are common in environmental data. Below is a brief overview of each model and the rationale for its selection.

**Random Forest Regressor**

Random Forest is an ensemble learning method based on decision trees. It builds multiple decision trees and merges their results to improve the accuracy of the model. By averaging the predictions from many trees, Random Forest helps reduce overfitting and improve model generalization. This model is particularly effective when capturing non-linear relationships and interactions between features. For this study**,** RandomForestRegressor from scikit-learn was used, and the model was evaluated using cross-validation to avoid overfitting and ensure robust performance.

**Gradient Boosting Regressor**

Gradient Boosting is another ensemble method that builds trees sequentially, each one correcting the residual errors of the previous tree. By focusing on residuals and iteratively improving the model, Gradient Boosting can create strong predictive models that perform well on complex datasets with non-linear patterns. In this study**,** GradientBoostingRegressor from scikit-learn was used to predict the AQI index.

**XGBoost Regressor**

XGBoost (Extreme Gradient Boosting) is a high-performance implementation of gradient boosting that includes additional features such as regularization to prevent overfitting and make the model more efficient. It is particularly known for its speed and effectiveness in handling large datasets. In this study, XGBRegressor from the XGBoost library was used.

***3.4 Model Training and Evaluation***

#### **Data Splitting**

To ensure robust model evaluation and prevent overfitting, the dataset was first split into two distinct subsets: **training** and **testing**. The data was split in an 80-20 ratio, with 80% of the data allocated for training the models and 20% reserved for testing and performance evaluation. The training set was used to fit the models, enabling them to learn the relationships between the input features (such as pollutant concentrations) and the target variable (AQI\_value). The testing set, which was kept separate from the training process, was then used to evaluate the model's ability to generalize to unseen data, providing a true estimate of its predictive performance.

**Cross-Validation**

#### To further enhance the reliability of model performance and reduce potential bias from a single train-test split, we implemented **k-fold cross-validation** with **k=10**. In this approach, the training data was randomly partitioned into 10 equal-sized subsets, or "folds." During each iteration of cross-validation, one fold was held out as the validation set, while the remaining nine folds were used for training. This process was repeated 10 times, with each fold serving as the validation set once, ensuring that all instances of the training data were used for both training and validation.

This technique provides a more comprehensive evaluation of the model by assessing its performance across different subsets of the data. The results from each fold are then averaged to produce a more robust estimate of the model's generalization ability. Cross-validation helps mitigate the risk of overfitting, as it ensures that the model's performance is consistent across different subsets of the data, rather than being overly optimized for a specific train-test split. By using k-fold cross-validation, we ensured that the model was evaluated on multiple different partitions of the data, providing a better understanding of how it would perform in real-world, unseen data scenarios.

### **Performance Evaluation Metrics**

After the completion of the cross-validation process, the performance of the models was assessed using key metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R²)**. These metrics provided insights into the model's accuracy and its ability to explain the variance in AQI values. For performance evaluation, we used three widely accepted metrics:

1. **Mean Absolute Error (MAE)** measures the average magnitude of the errors in predictions, without considering their direction. It calculates the average of the absolute differences between the predicted and actual values. MAE is a straightforward metric, providing an intuitive understanding of the model's overall accuracy by indicating how far, on average, the predictions are from the actual values.
2. **Root Mean Squared Error (RMSE)** is similar to MAE but gives more weight to larger errors due to squaring the differences. RMSE provides a clearer measure of how well the model fits the data, especially when large errors are undesirable. It is sensitive to outliers, making it useful for identifying models that fail to predict extreme values accurately.
3. **R-squared (R²)** is a statistical measure of how well the model's predictions approximate the actual data. It indicates the proportion of variance in the dependent variable that is predictable from the independent variables. R² ranges from 0 to 1, where a value closer to 1 indicates that the model explains most of the variance in the data, and a value closer to 0 suggests poor predictive power.

***3.5 Software and Tools***

This research utilized a combination of tools and technologies for data preprocessing, modeling, and analysis. Initially, **Turbo Prep in RapidMiner AI Studio** was employed for efficient and intuitive preprocessing tasks. Subsequently, the project transitioned to **Python (version 3.12)** due to its flexibility and rich ecosystem of libraries tailored for machine learning and data science. Key Python libraries included **pandas** and **NumPy** for data manipulation, **Matplotlib** and **Seaborn** for visualization, and scikit-learn for implementing regression models and computing performance metrics. Advanced boosting techniques were implemented using **XGBoost** and **GradientBoostingRegressor**, which offered superior performance and scalability. Development and experimentation were conducted in **Jupyter Notebook**, an interactive environment that allowed iterative coding and visualization.

1. **RESULT AND DISCUSSION**

In this section, we present and discuss the empirical results obtained from training and evaluating the three models: Random Forest, Gradient Boosting, and XGBoost. The performance metrics used were Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). These metrics provide a comprehensive evaluation of the models in terms of prediction accuracy and generalization ability.

Table 4.1. Model Performance Overview

| **Model** | **MAE** | **RMSE** | **R²** |
| --- | --- | --- | --- |
| Random Forest | 15.09 | 22.91 | 0.9371 |
|  |  |  |  |
| Gradient Boosting  XGBoost | 14.10  13.91 | 22.20  21.31 | 0.9388  0.9436 |

#### From the results, **XGBoost** achieved the highest R² score (0.9436), indicating its superior ability to explain the variance in AQI values. **Gradient Boosting** followed closely with an R² score of 0.9388, while **Random Forest** had a slightly lower R² (0.9371). In terms of MAE and RMSE, XGBoost outperformed the other models, achieving the lowest values (MAE: 13.91, RMSE: 21.31), reflecting its ability to minimize prediction errors more effectively. These results position XGBoost as the most accurate model for predicting AQI values in this study.

#### ***Comparative Analysis***

* **XGBoost** demonstrated a robust ability to generalize, as evidenced by its lowest RMSE (21.31) and MAE (13.91), indicating fewer large prediction errors and more precise predictions overall. Its ensemble nature and regularization capabilities allow it to capture complex relationships effectively, making it the best-performing model.
* **Gradient Boosting**, while slightly behind XGBoost in terms of R² (0.9388), showed competitive performance with a relatively low RMSE (22.20) and MAE (14.10). Its sequential approach to minimizing residual errors contributed to reasonable predictions but indicated some room for improvement.
* **Random Forest**, with an R² of 0.9371, also delivered strong results, particularly excelling in its ability to explain data variability. However, with slightly higher MAE (15.09) and RMSE (22.91) compared to XGBoost, it ranked second overall in this study.

***4.2 Visual Comparison of Predicted vs. Actual AQI***

To evaluate the predictive accuracy of the models, we compared the actual AQI values with the predicted values using a scatter plot. This visualization helps assess how closely the models' predictions align with the observed data. Ideally, points should lie along the diagonal line (y = x), indicating perfect predictions. Deviations from this line represent prediction errors, with smaller deviations reflecting higher accuracy. The plot highlights the relative performance of Random Forest, Gradient Boosting, and XGBoost in capturing the true AQI values.

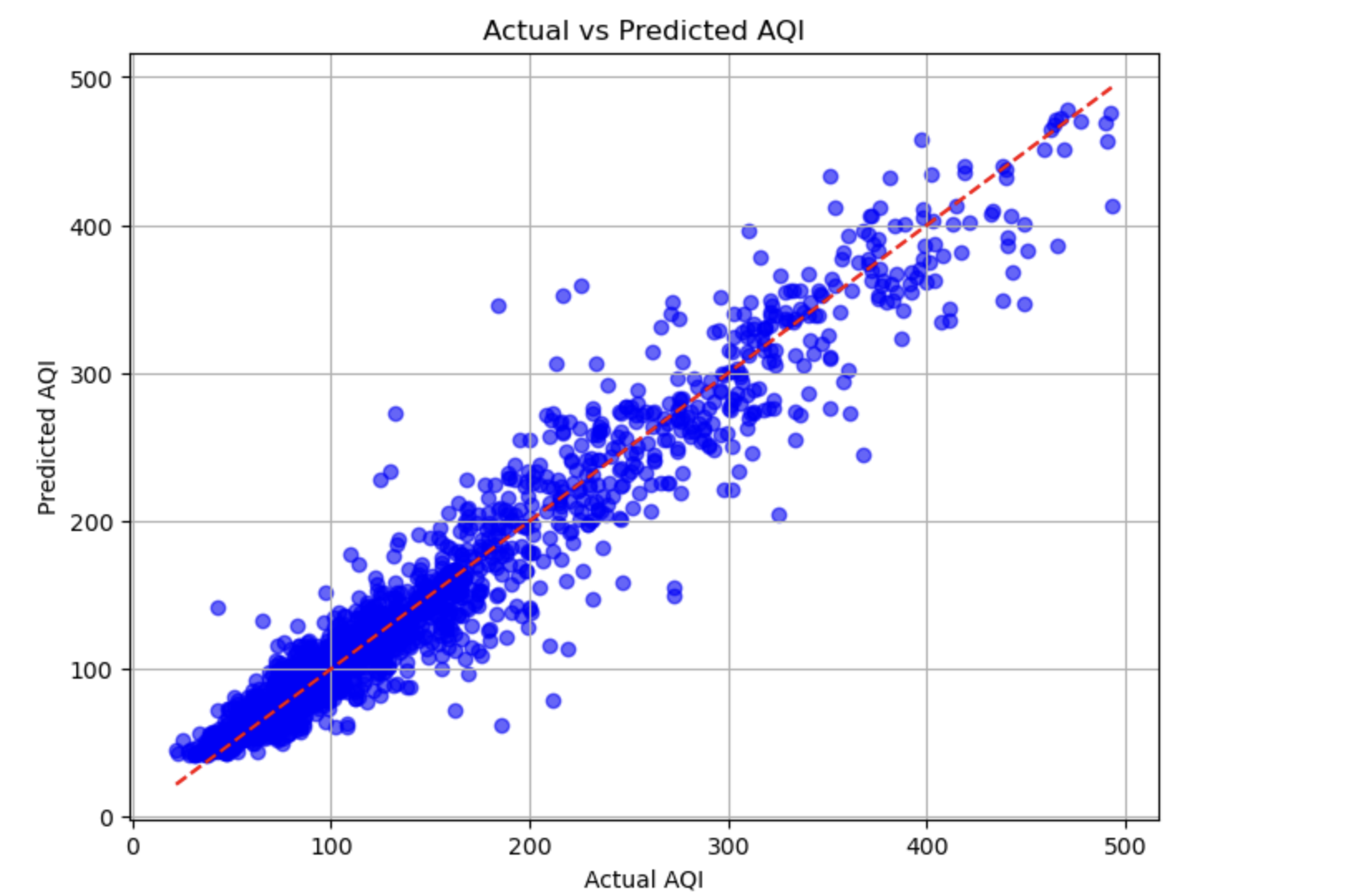
****

Fig 4.2: Random Forest

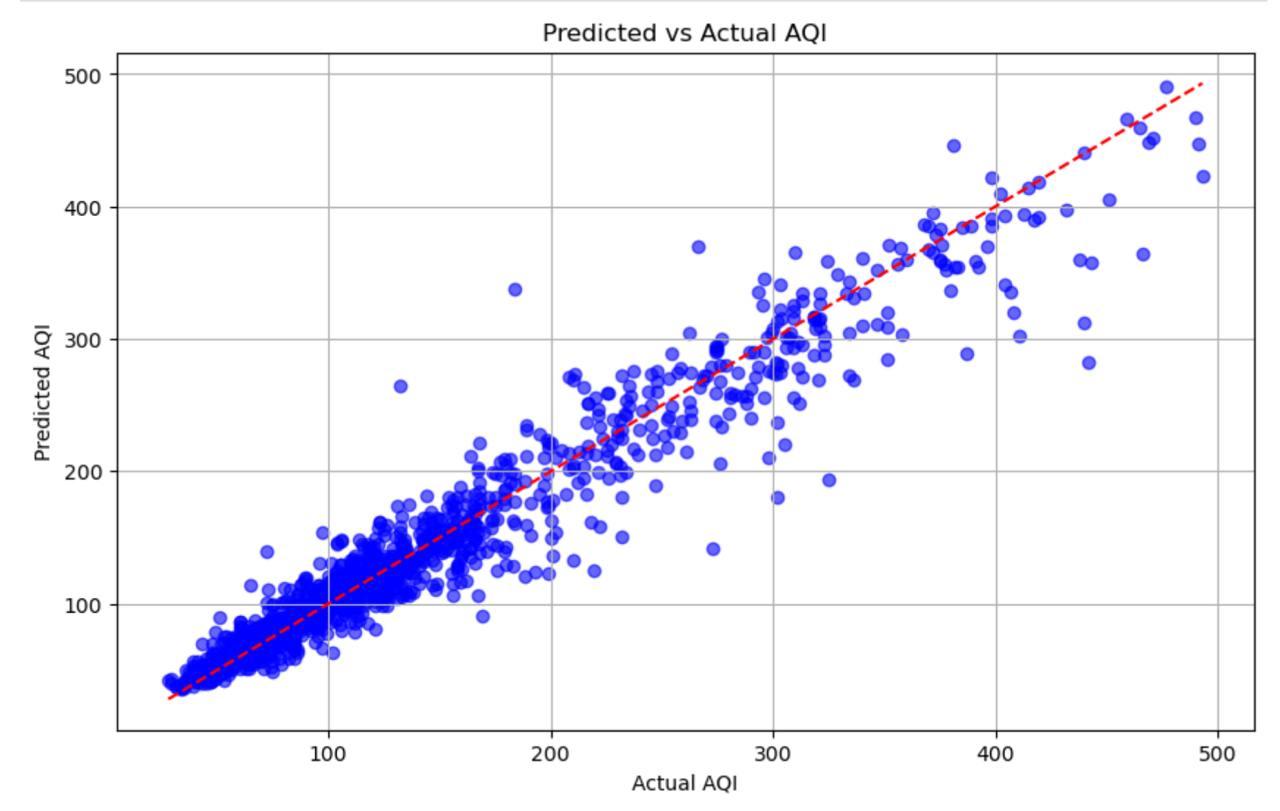


Fig 4.3: Gradient Boosting

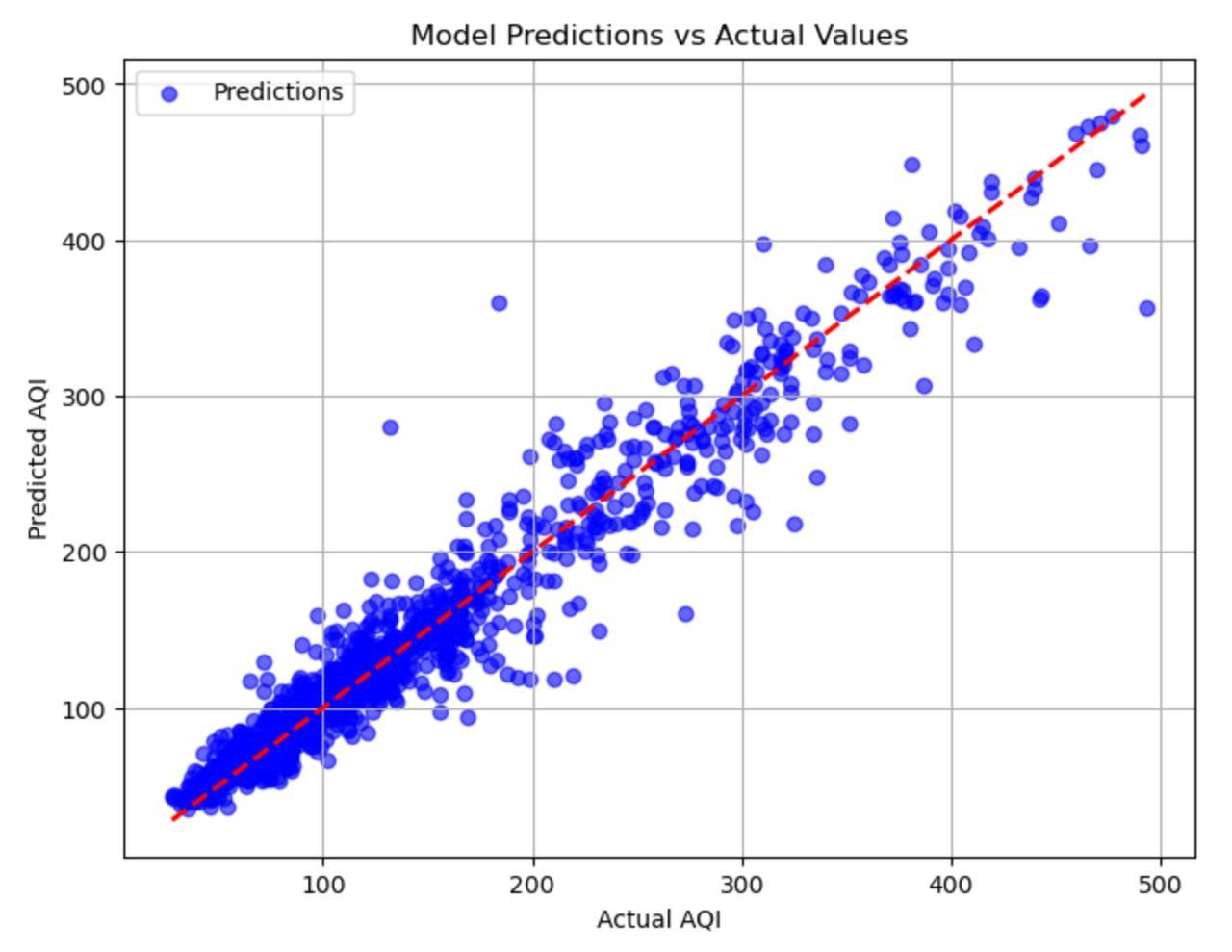
****

Fig 4.4: XGBoost

***4.3 Feature Importance Analysis***

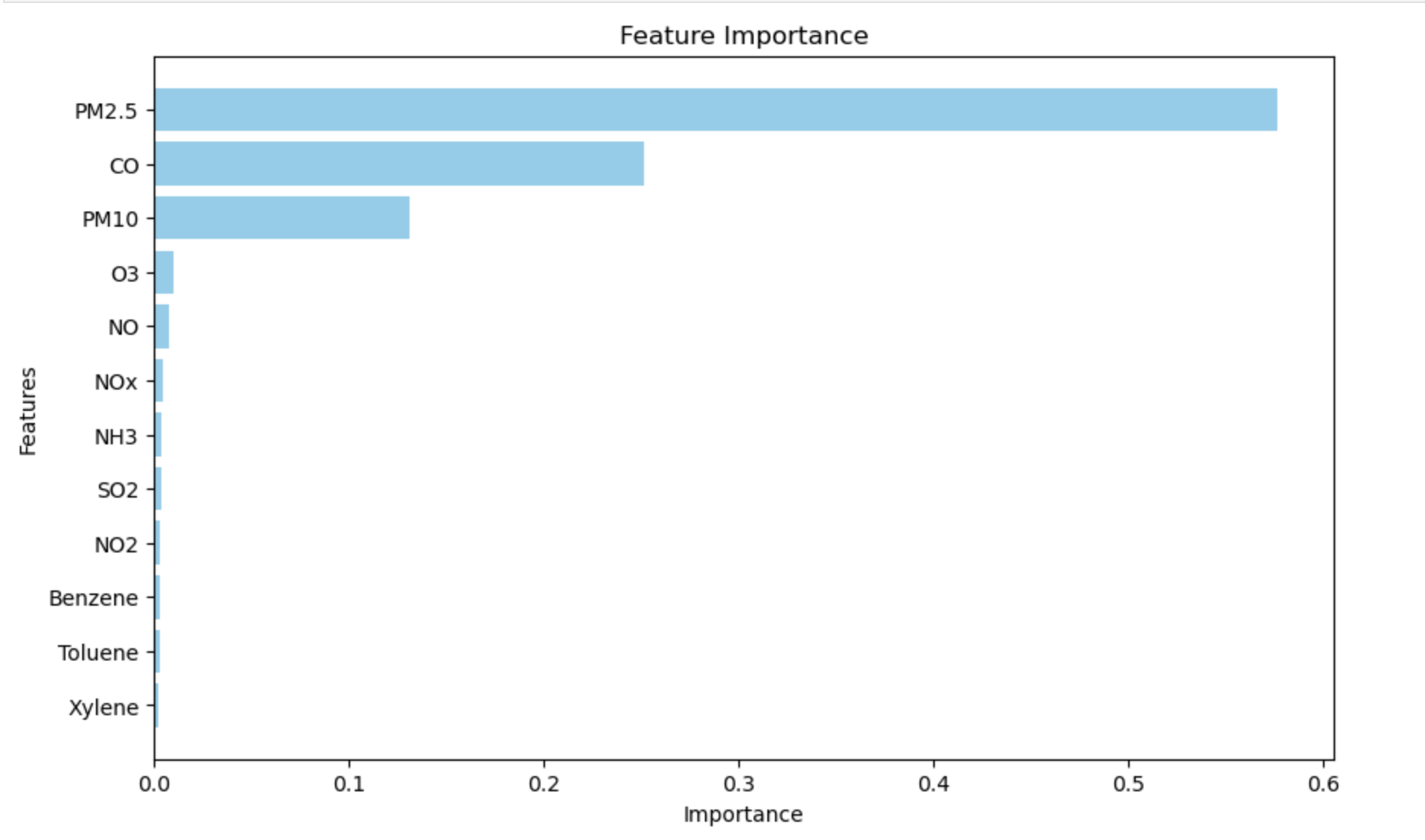


Fig 4.5: Feature Importance Graph

***4.4 RMSE using Cross-Validation***

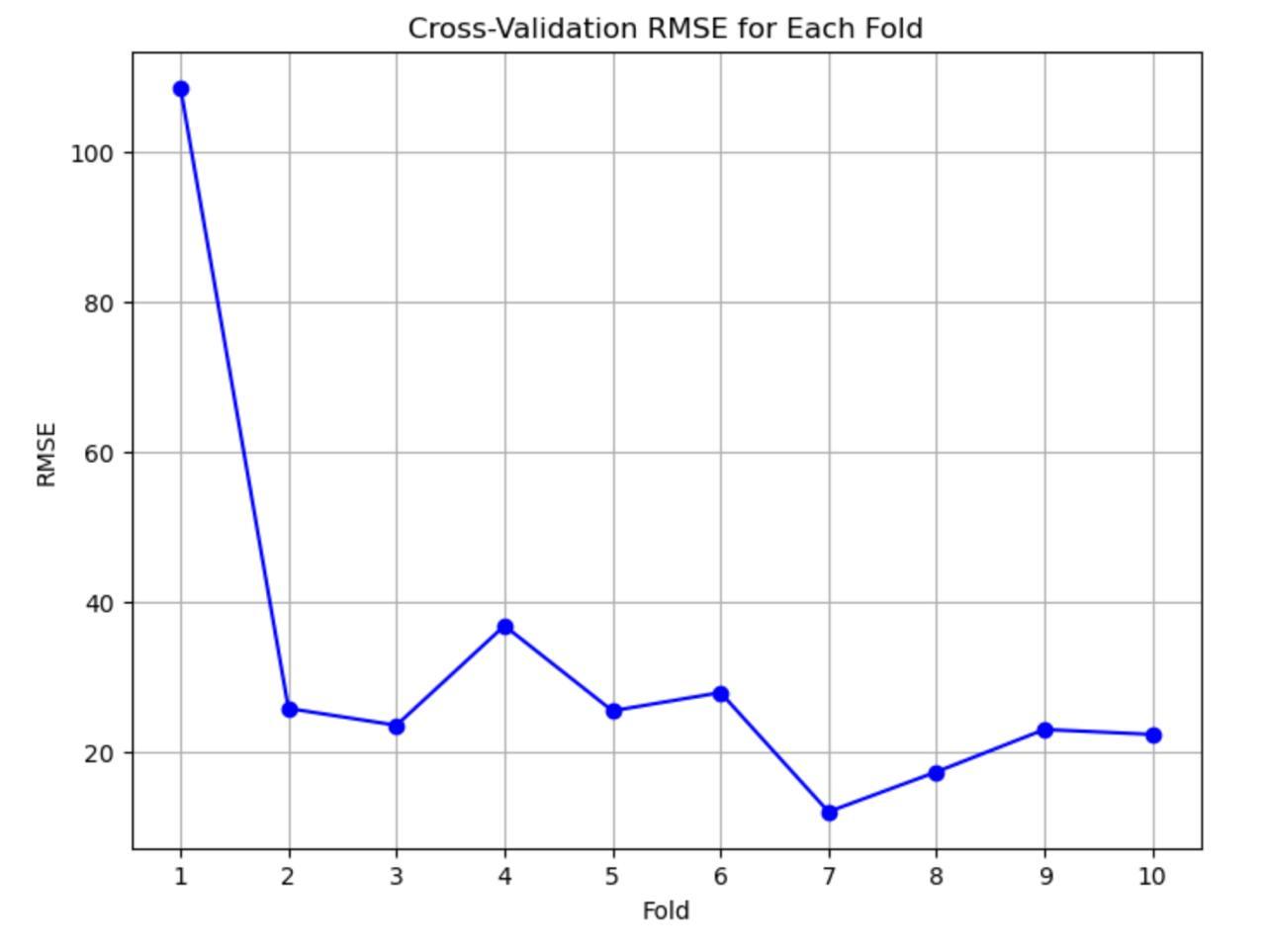
******

Fig 4.6: Cross\_validation RMSE for Each fold

***4.5 Model Metrics Comparison***

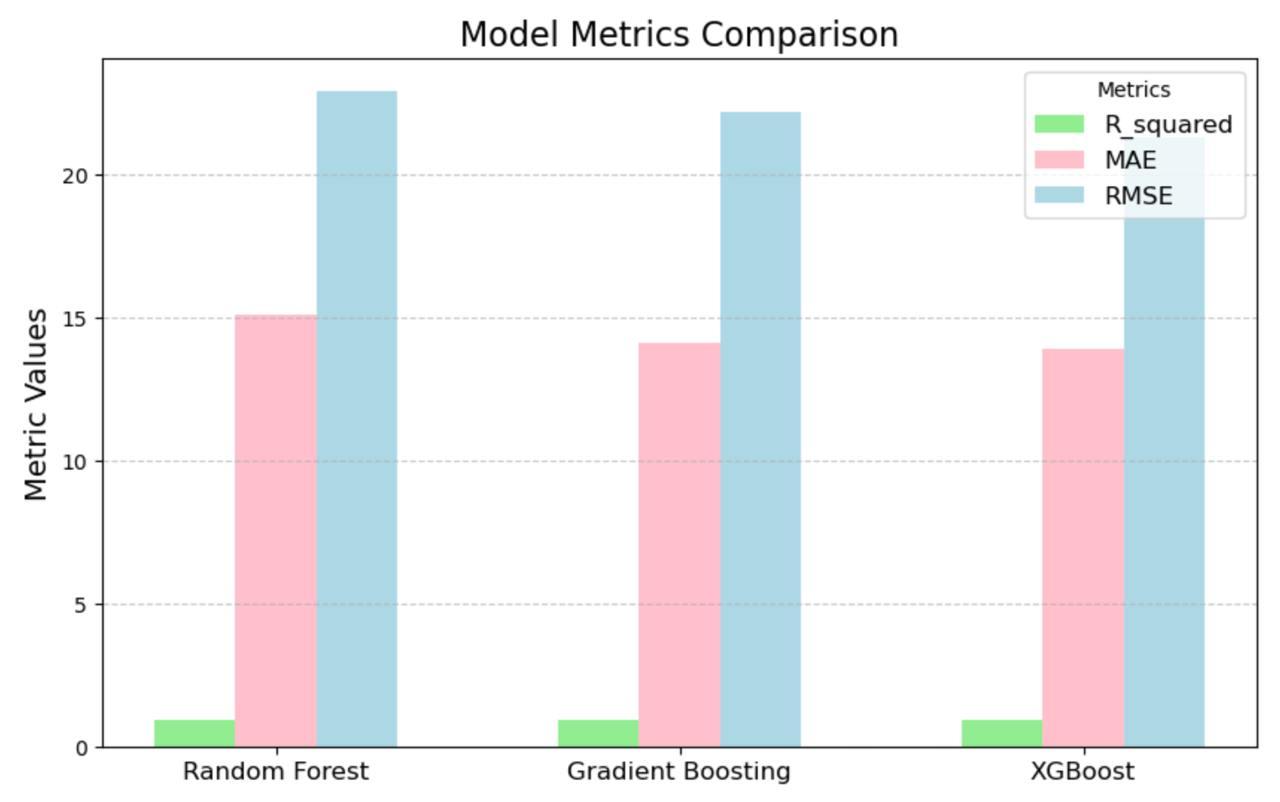
******

Fig 4.7: Comparison of Model Metrics

The feature importance analysis revealed that PM2.5, CO and PM10 were the most significant contributors to AQI predictions, aligning with domain knowledge about their critical impact on air quality.

Overall, the results demonstrate that Gradient Boosting and XGBoost are well-suited for predicting AQI with high accuracy and reliability. These models could be deployed for real-time air quality monitoring, helping policymakers and urban planners implement effective measures to mitigate pollution.

5. **CONCLUSION**

This study presents a comprehensive analysis of air quality prediction using three advanced machine learning models: Random Forest, Gradient Boosting, and XGBoost. Leveraging publicly available air quality data from India (2015–2020), we preprocessed, imputed, and engineered features to create a robust dataset suitable for predictive modeling. The models were trained and evaluated using an 80-20 train-test split combined with 10-fold cross-validation to ensure consistency and generalizability.

Among the models, XGBoost achieved the best performance with the lowest RMSE (21.31), MAE (13.91), and the highest R² (0.9436), demonstrating its ability to capture complex patterns. Gradient Boosting closely followed with an RMSE of 22.20 and R² of 0.9388, while Random Forest, with an RMSE of 22.91 and R² of 0.9371, still provided competitive predictions due to its ensemble robustness. The evaluation metrics—MAE, RMSE, and R²—highlighted the strengths and limitations of each model. Random Forest excelled due to its ensemble nature and resilience against overfitting, making it particularly suitable for tasks with non-linear relationships and interactions. XGBoost’s strong performance underscores its capacity to balance bias-variance tradeoffs, while Gradient Boosting’s results reflect its ability to build strong predictive models through residual correction.

This study also emphasizes the importance of rigorous validation techniques such as k-fold cross-validation, which provided a robust assessment of model reliability across different data subsets. Such practices are crucial for ensuring the models' practical applicability to real-world scenarios. From a broader perspective, the findings underscore the potential of machine learning models in urban air quality monitoring and prediction. These models can provide actionable insights for policymakers, enabling targeted interventions to mitigate pollution and protect public health. However, certain limitations should be noted. The models rely heavily on the quality and comprehensiveness of input data. Factors such as regional variability, data sparsity in smaller cities, and the exclusion of meteorological parameters in this study may affect the models' broader applicability.

Future research should explore integrating additional environmental and meteorological factors, temporal trends, and advanced deep learning architectures for further improving prediction accuracy. Additionally, deploying these models in real-time air quality monitoring systems could provide dynamic, actionable insights to address urban pollution challenges effectively.

In conclusion, this research demonstrates that machine learning models, particularly ensemble methods like Random Forest and XGBoost, are powerful tools for air quality prediction. While each model has its strengths, the findings highlight the importance of tailoring model selection and preprocessing techniques to the specific characteristics of the data and prediction objectives. This work sets a foundation for further exploration and application of AI-driven solutions in environmental monitoring.

**Acknowledgement**

We would like to express our deepest gratitude to our **Asst.Prof.Dr. Kritsada Sriphaew,** Dean of Rangsit International College, and our teaching assistant, **Ms. Aye Khin Khin Hphone**, for their exceptional guidance and insightful feedback. Our sincere appreciation extends to the various online resources, forums, and communities that provided us with access to essential tools and tutorials. Finally, I am grateful to my team members for their collaboration, enthusiasm, and hard work. Their contributions to data collection, model development, and analysis were integral to the success of this research. Working together has been an enriching experience, and I look forward to future collaborations.

**References**

* RapidMiner AI Studio. (n.d.). Retrieved November 7, 2024, from <https://www.rapidminer.com>
* Central Pollution Control Board. (n.d.). *Air quality data for cities across India*. Retrieved November 7, 2024, from <https://www.cpcb.nic.in>
* Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. *Complexity*, *2020*, 1–23. <https://doi.org/10.1155/2020/8049504>
* IEEE Hong Kong Section 50th Anniversary 1972-2022: Advance Technology for Huminity – The Tech-Biz Intelligence. (2022). *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON)*, 1–35. <https://doi.org/10.1109/tencon55691.2022.9977897>
* Dun, M., Xu, Z., Chen, Y., & Wu, L. (2020). Short-Term air quality prediction based on fractional grey linear regression and support vector machine. *Mathematical Problems in Engineering*, *2020*, 1–13. <https://doi.org/10.1155/2020/8914501>
* Meo, S. A., Almutairi, F. J., Abukhalaf, A. A., Alessa, O. M., Al-Khlaiwi, T., & Meo, A. S. (2021). Sandstorm and its effect on particulate matter PM 2.5, carbon monoxide, nitrogen dioxide, ozone pollutants and SARS-CoV-2 cases and deaths. *The Science of the Total Environment*, *795*, 148764[. https://doi.org/10.1016/j.scitotenv.2021.148764](about:blank)
* Singh, R., Raghav, S., Maini, T., Singh, M., & Arquam, M. (2022). *Air quality prediction using machine learning*. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.4157651>
* *Cross-Validation - Amazon Machine Learning*. (n.d.). <https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html>
* Yu, R., Yang, Y., Yang, L., Han, G., & Oguti, A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors, 16*(1), 86. <https://doi.org/10.3390/s16010086>