iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XII **Month of publication:** December 2024

DOI: https://doi.org/10.22214/ijraset.2024.66036

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XII Dec 2024- Available at www.ijraset.com

Predictive Modelling for Air Quality: A Multi-Model Regression Approach

Dr. Benakappa S M[1], Trupti K N[2], Sindhu G B[3] , Thejaswini H[4] , Varun S Sindhe[5]
Department of CS&E, JNN College of Engineering, Shivamogga, Karnataka, India

Abstract: ***Air pollution poses a significant threat, necessitating reliable predictive models for effective air quality management. This study employs a multi-model regression approach for AQI prediction using techniques like Extra Trees Regressor, SVR, LightGBM, and CatBoost. Real-time datasets from CPCB, combined with meteorological and pollutant data, are pre-processed to handle missing values, outliers, and imbalances. Feature selection enhances accuracy by focusing on significant variables. The performance evaluation of the models revealed that the ExtraTrees Regressor was the best performer, achieving the highest R^2 score of 0.8660 and the lowest RMSE of 56.98. These results highlight ExtraTrees Regressor's superior ability to explain AQI variance and minimize prediction errors, making it the most effective model for AQI forecasting in this study. Visualization techniques compare model outcomes, while distributed computing ensures real-time AQI forecasting, supporting sustainable urban air quality strategies.***

Keywords: Air Quality Index (AQI), Feature Selection, Root Mean Square Error (RMSE), R^2 Score, Environmental Sustainability.

I. INTRODUCTION

Air pollution remains one of the most critical challenges faced by urban and rural regions globally, with profound effects on human health, ecosystems, and climate change. The rapid growth of industrial activities, increasing vehicular emissions, and urbanization have exacerbated pollution levels in many cities, resulting in a marked decline in air quality [1]. According to the World Health Organization (WHO), air pollution is a leading environmental risk factor, causing millions of premature deaths each year. Long-term exposure to polluted air can lead to respiratory diseases, cardiovascular conditions, and finally premature mortality. Furthermore, pollutants like particulate matter ($PM_{2.5}$, PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) can degrade environmental quality, harm ecosystems, and contribute to global warming. Therefore, predicting air quality accurately is essential for reducing exposure to harmful pollutants and mitigating their long-term impacts on human health and the environment [2]. The Air Quality Index (AQI) is a standardized tool used worldwide to quantify and communicate the concentration of air pollutants. AQI values range from 0 to 500, with higher values indicating poorer air quality. This index helps authorities issue timely warnings and guidelines to the public, particularly vulnerable populations such as children, the elderly, and individuals with pre-existing health conditions. While AQI provides a clear snapshot of air quality, its prediction is far more complex. Accurate forecasting of AQI is necessary for pre-emptive public health interventions, urban planning, and environmental management. However, air quality prediction is a challenging task due to the complexity of air pollution dynamics and the influence of multiple factors. Historically, statistical models have been used to forecast air quality, but these methods often struggle to capture the complex, non-linear relationships between air pollutants and meteorological variables such as temperature, humidity, wind speed, and atmospheric pressure. Moreover, traditional models may fail to accommodate missing data, measurement errors, and changes in pollutant levels due to local factors. With the advancements in machine learning and data science, more sophisticated approaches have been developed that can overcome these limitations by learning patterns from large, complex datasets [3].

Machine learning (ML) models, especially regression techniques, have shown great promise in predicting AQI by identifying patterns in historical and real-time data. These models can learn from various data sources, including air quality monitoring stations, meteorological data, and satellite observations, to provide accurate and dynamic predictions. One of the significant advantages of machine learning models is their ability to handle high-dimensional data, account for complex relationships between multiple pollutants, and provide real-time forecasts. In addition, these models can adapt to new data, improving their predictive performance over time. The primary objective of this study is to apply multi-model regression approaches to predict AQI using various machine learning techniques. This approach combines multiple algorithms to enhance prediction accuracy by selecting the best-performing model based on specific evaluation criteria. Techniques such as Extra Trees Regressor, Support Vector Regressor (SVR), LightGBM, and CatBoost are explored to evaluate their effectiveness in predicting AQI in different environmental conditions.

These models were chosen because they excel in handling complex, non-linear relationships and large datasets, which are common in environmental studies.

The research focuses on several key aspects of AQI prediction. First, the study utilizes real-time data obtained from the Central Pollution Control Board (CPCB) and other sources to create a comprehensive dataset that includes pollutants such as $PM_{2.5}$, NO_2, SO_2, and O_3, alongside meteorological data like wind speed, temperature, and humidity. The preprocessing techniques such as missing data handling, outlier detection, and feature scaling are applied to ensure that the models receive clean and reliable data for training and evaluation.

Second, the research explores various feature engineering techniques to enhance model performance. By selecting the most relevant features and creating new ones (e.g., pollutant ratios, time-lagged features), the models can capture important trends and relationships that may not be immediately apparent in the raw data [4]. This step is crucial in improving the accuracy and interpretability of the prediction models.

Third, the study evaluates the performance of the models using standard metrics like R^2 (coefficient of determination) and Root Mean Square Error (RMSE). R^2 measures how well the model explains the variance in the AQI, while RMSE assesses the magnitude of errors in prediction. Both of these metrics are critical for understanding how well the models predict real-world AQI values and provide insights into their generalization ability. Moreover, to visualize and compare the models' performance, scatter plots and bar charts are employed. These visualizations allow for clear comparison between actual and predicted AQI values, offering intuitive insights into how well each model captures air quality trends.

II. LITERAUTURE SURVEY

In this section, various authors have presented various air quality prediction using diverse models and methodologies.

In [5], machine learning models—Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression—were used to predict AQI in four Indian cities. Data balancing through SMOTE improved accuracy, with RFR and CatBoost emerging as the top-performing models. However, the study's limited geographic scope restricted its generalizability.

In [6], applied KNN, Gaussian Naive Bayes, SVM, Random Forest, and XGBoost to AQI data from 23 Indian cities. XGBoost demonstrated the highest accuracy, highlighting its scalability for large datasets. However, variations in model performance across cities revealed challenges in adapting models to diverse regional conditions.

In [7], SARIMA, SVM with RBF kernel, and LSTM models were compared for AQI prediction in Ahmedabad. SVM with RBF kernel achieved the highest accuracy, demonstrating its effectiveness for industrial regions, though the study was limited to a single city.

In [8], the ARIMA model to predict pollutant levels in Surat. While the model effectively forecasted short-term AQI trends and analysed lockdown-related pollution reductions, the lack of alternative model comparisons limited its insights.

In [9], evaluated CatBoost, Random Forest, and XGBoost models for AQI prediction in Visakhapatnam. CatBoost achieved the best performance ($R^2 = 0.9998$) and identified $PM_{2.5}$ and PM_{10} as key pollutants. However, the study's scope was restricted to a single city without real-time testing.

In [10], high-resolution spatiotemporal models for six pollutants were developed using satellite and ground monitoring data. The study provided detailed AQI maps for Shanghai, identifying O_3 as a dominant pollutant. However, the exclusion of meteorological data limited the model's accuracy and broader applicability.

In [11], employed GIS-based spatial interpolation and Geographically Weighted Regression models to predict AQI in Iraq using remote sensing and ground station data. The study highlighted $PM_{2.5}$ as a major pollutant but focused on a single season and lacked real-time monitoring, restricting its generalizability.

In [12], developed a linear regression model in R to predict pollutant levels in Belgrade, focusing on $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO. The approach was cost-effective and adaptable, but it was limited to winter data and lacked meteorological integration, impacting its prediction precision.

In [13], a hybrid VMD–EEMD–LSTM model was proposed for ozone prediction. The model achieved high accuracy ($R^2 = 98\%$) by handling complex, non-stationary data but focused solely on ozone and did not incorporate meteorological factors.

In [14], introduced a multi-scale, attention-enhanced CNN model combining low-cost sensor data, satellite AOD, and LiDAR-derived 3D urban features for $PM_{2.5}$ prediction at 30-m resolution. The model effectively captured fine-scale pollution patterns and urban-rural disparities but was limited to Denton County, Texas, requiring broader validation for widespread application.

Table 1. Taxonomy of AQI prediction.

Ref Pap No.	Methodology	Contributions	Advantages	Limitations
N. Srinivasa Gupta et al. (2023) [5]	Applied SVR, RFR, and CatBoost for AQI prediction in four Indian cities with SMOTE for data balancing.	Identified RFR and CatBoost as top-performing models for AQI prediction.	Improved accuracy through balanced data.	Limited to four cities, reducing generalizability.
K. Kumar & B.P. Pande (2023) [6]	Used KNN, Naive Bayes, SVM, Random Forest, and XGBoost on six years of AQI data from 23 Indian cities.	Found XGBoost to have the highest accuracy among tested models.	Scalable framework for large datasets.	Variations in model performance across cities limit adaptability.
Nilesh N. Maltare & Safvan Vahora (2023) [7]	Compared SARIMA, SVM with RBF kernel, and LSTM for AQI prediction in Ahmedabad.	Demonstrated SVM with RBF kernel's high accuracy in AQI prediction.	Effective for industrial regions.	Focused on Ahmedabad; limited testing in other geographies.
H. N. Mahendra et al. (2023) [8]	Applied ARIMA model to predict pollutant levels in Surat using COVID-19 lockdown data.	Showcased ARIMA's capability for short-term AQI forecasting.	Useful for analysing lockdown effects on air quality.	Did not explore alternative models, limiting comparative insights.
G. Ravindiran et al. (2023) [9]	Evaluated CatBoost, Random Forest, XGBoost, and others for AQI prediction in Visakhapatnam.	CatBoost outperformed other models (R^2 = 0.9998); identified $PM_{2.5}$, PM_{10} as key contributors.	Highly accurate AQI predictions.	Restricted to one city, no real-time applicability.
Yiyi Wang et al. (2023) [10]	Developed high-resolution spatiotemporal models for six pollutants in Shanghai using satellite and ground data.	Created detailed AQI maps revealing spatial variability and key pollutants.	High-resolution analysis for urban pollution.	Excluded meteorological factors like wind and humidity.
Huda Jamal Jumaah et al. (2023) [11]	Used GIS, Least Squares, and GWR to predict AQI in Iraq using remote sensing and ground station data.	Generated AQI maps identifying $PM_{2.5}$ as a major pollutant.	Effective framework for under-monitored regions.	Focused on a single season; lacked real-time monitoring.
Zoltan Kazi, Snezana Filip, and Ljubica Kazi (2024) [12]	Built a linear regression model in R for pollutant prediction in Belgrade.	Provided a cost-effective pollutant prediction tool for $PM_{2.5}$, PM_{10}, SO_2, and others.	Adaptable and simple to use.	Limited to winter data; no meteorological factors included.
Tang et al. (2024) [13]	Proposed hybrid VMD–EEMD–LSTM model for ozone prediction using dual series decomposition.	Achieved high accuracy (R^2 = 98%), outperforming traditional models.	Effective for complex, non-stationary data forecasting.	Focused only on ozone; lacked meteorological integration.
Lu Liang et al. (2024) [14]	Developed multi-scale CNN using sensor, satellite, and LiDAR data for $PM_{2.5}$ prediction at 30-m resolution.	Highlighted urban-rural pollution disparities with high accuracy (R^2 = 0.80).	Captures fine-scale pollution variations.	Limited to Denton County; needs broader validation.

III. AIR QUALITY PREDICTION APPROACH

The air quality prediction process involves several sequential steps, starting from data acquisition to model evaluation. The primary objective is to predict the Air Quality Index (AQI) by leveraging machine learning algorithms with high accuracy and reliability. The approach ensures the inclusion of vital preprocessing techniques to prepare raw data, followed by the implementation of various regression models for prediction and performance evaluation.

The provided flowchart visually represents the proposed methodology for predicting air quality. It outlines a systematic workflow starting with loading datasets and installing necessary libraries, followed by preprocessing steps such as handling missing values and normalizing the data. The dataset is then split into training and testing sets, and scaling is applied to standardize the features. Once the data is prepared, the training of machine learning models is carried out, followed by their evaluation based on performance metrics such as R^2 and Root Mean Square Error (RMSE). A decision step evaluates the model's accuracy, and adjustments are made to parameters or models if required. Finally, the model with the best performance is selected for predicting AQI.

This step-by-step approach ensures a comprehensive analysis of data and optimizes the prediction process. The figure 1 gives a detailed explanation of the components in the flowchart, which serves as the foundation of the air quality prediction methodology.

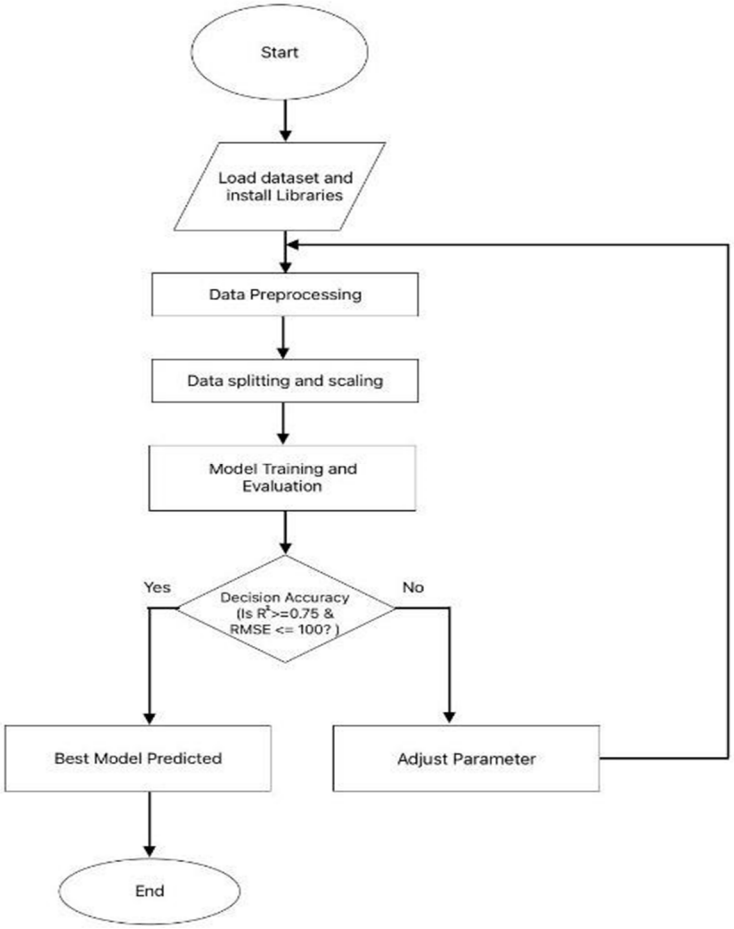


Figure 1. Flowchart of prediction model of AQI.

A. Dataset Description

The dataset utilized for this project comprises air quality data collected from the city of Amaravati. It records various pollutants and their concentrations on specific dates, along with an overall assessment of air quality. The dataset includes key features such as the date of data collection, pollutant concentrations, and the corresponding Air Quality Index (AQI) values. Each entry in the dataset represents measurements taken for multiple pollutants, including Particulate Matter ($PM_{2.5}$ and PM_{10}), Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Ammonia (NH_3), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), and Ozone (O_3). Additionally, it tracks concentrations of volatile organic compounds like Benzene (C_6H_6), Toluene (C_7H_8), and Xylene (C_8H_{10}).

The AQI, an essential feature of the dataset, is calculated based on pollutant concentrations and is used to categorize the air quality into qualitative buckets such as "Good," "Moderate," or "Poor." For example, in this dataset, a recorded AQI of 184 falls into the "Moderate" category, indicating a noticeable level of pollution. The dataset's structure allows for a comprehensive analysis of the relationship between different pollutants and their contribution to air quality.

This rich collection of variables not only facilitates the training and evaluation of machine learning models but also enables deeper insights into the dynamics of air quality. The dataset's temporal dimension (i.e., measurements taken on specific dates) further allows for the exploration of trends and patterns over time, making it an invaluable resource for predicting AQI and aiding in environment.

B. Data Pre-Processing

Data pre-processing is a crucial step in the air quality prediction approach as it ensures that the dataset is clean, consistent, and suitable for machine learning analysis. The dataset undergoes several pre-processing techniques to handle issues such as missing values, outliers, and scaling inconsistencies, which can impact the accuracy of the predictive models.

The first step in data pre-processing involves addressing missing values, which are common in real-world datasets. Missing data points are either filled using statistical methods such as mean, median, or mode imputation or removed if they are minimal and their absence does not significantly affect the overall data structure. This ensures that the dataset remains complete and representative of the underlying air quality patterns. Outlier detection and handling are also performed to ensure that extreme values, which may result from measurement errors or anomalies, do not skew the predictions. Outliers are either corrected or removed based on their impact on the dataset. Next, categorical variables, such as the AQI bucket (e.g., "Good," "Moderate," "Poor"), are encoded into numerical formats to make them compatible with machine learning models. Techniques like label encoding or one-hot encoding are used for this purpose. Normalization or scaling of numerical features is then performed to ensure that all variables contribute equally to the model's performance. Pollutant concentrations often have varying units and ranges, which can lead to biased predictions if not scaled. Standardization methods, such as Min-Max scaling or Z-score normalization, are applied to bring all features into a uniform range. Lastly, the dataset is split into training and testing sets. The training set is used to train the machine learning models, while the testing set is reserved for evaluating the models' performance. This split is essential for assessing the models' generalization capabilities on unseen data. Through these data pre-processing steps, the dataset is transformed into a format that is optimal for machine learning analysis, enhancing the accuracy and reliability of the air quality prediction models.

C. AQI Calculation

The Air Quality Index (AQI) is a critical measure used to quantify air quality and assess the impact of pollution on human health and the environment. It aggregates the concentrations of various pollutants, such as PM2.5, PM10, NO2, SO2, O3, and others, into a single, standardized value. AQI is categorized into qualitative buckets, such as "Good," "Moderate," or "Poor," to provide an easy-to-understand representation of air quality.

In the model, the AQI calculation involves determining the AQI for individual pollutants based on their concentrations. The calculation follows standards set by environmental agencies, where the pollutant with the highest AQI value determines the overall AQI. The formula for calculating the AQI for a specific pollutant is:

$$\text{AQI}=\frac{I_{high-}\ I_{low}}{C_{high-}\ C_{low}}\ x\ (C\ -\ C_{low})\ +\ I_{low} \quad (1)$$

C: The concentration of the pollutant (e.g., $PM_{2.5}$, PM_{10}, NO_2) measured from monitoring stations.
C_low: The concentration breakpoint that is less than or equal to C.
C_high: The concentration breakpoint that is greater than or equal to C.
I_low: The AQI value corresponding to C_low.
I_high: The AQI value corresponding to C_high.

D. Feature Engineering

Feature engineering plays a vital role in improving the accuracy and efficiency of AQI prediction models by transforming raw data into meaningful inputs. The focus is on selecting key variables such as $PM_{2.5}$, PM_{10}, NO_2, and O_3, along with meteorological factors like temperature and humidity, which directly influence air quality. Unnecessary features are excluded to reduce noise and complexity.

Derived features are created to enhance model understanding, including pollutant ratios (e.g., $PM_{2.5}/PM_{10}$) and temporal features like the previous day's AQI. Feature scaling, such as Min-Max normalization, ensures all variables have a uniform range, preventing any single feature from dominating the model training process. Categorical features like AQI buckets are numerically encoded to make them compatible with machine learning algorithms. These engineering steps ensure that the data is optimized for modeling, enabling more accurate and reliable AQI predictions.

E. Splitting Data

Splitting the dataset into training and testing sets is a crucial step in building and evaluating machine learning models for AQI prediction. This ensures that the model learns patterns from one portion of the data (training set) and is validated on an unseen portion (testing set) to assess its performance and generalization capability.

In this model, the dataset is split into two subsets: the training set, which comprises 80% of the data, and the testing set, which includes the remaining 20%. The training set is used to fit the machine learning models, allowing them to learn the relationships between input features (pollutant concentrations and meteorological factors) and the target variable (AQI). The testing set is reserved for evaluating the model's predictive accuracy and robustness on unseen data, ensuring it performs well on real-world scenarios.

This split is performed randomly to ensure that the subsets represent the overall distribution of the data. Randomization minimizes bias and prevents overfitting, where the model might memorize specific data points instead of learning general patterns. By maintaining an appropriate balance between training and testing data, the approach ensures that the models are both accurate and capable of generalizing to new datasets.

F. Regression Models Construction

The final step in the air quality prediction approach involves constructing regression models to predict the Air Quality Index (AQI). For this task, the models are trained using advanced regression techniques, namely Extra Trees Regressor, Support Vector Regressor (SVR), LightGBM, and CatBoost. Each capable of capturing the complex relationships between pollutant concentrations, meteorological variables, and AQI. The following models are implemented and analyzed.

1) *Extra Trees Regressor:* This is an ensemble learning method based on Decision Trees. Extra Trees Regressor improves prediction accuracy by constructing multiple Decision Trees during training and averaging their predictions. Unlike Random Forest, it splits nodes randomly, which reduces variance and enhances computational efficiency. This model is particularly effective for handling high-dimensional datasets with complex, non-linear relationships between input features and AQI. It leverages historical air pollution data to identify patterns and make accurate AQI predictions.
2) *Support Vector Regressor (SVR):* SVR is a supervised machine learning algorithm that utilizes hyperplanes to model the relationship between input features and target variables. By using a kernel trick, SVR can map the data into higher-dimensional spaces to capture non-linear patterns in pollutant concentrations and AQI. The model optimizes the margin of error between predicted and actual values, making it suitable for precise AQI forecasting. SVR is especially useful when the dataset contains noise or outliers, as it focuses on minimizing prediction errors while maintaining model robustness.
3) *CatBoost Regressor:* CatBoost is a gradient boosting algorithm specifically designed to handle categorical and numerical data efficiently. It reduces the complexity of handling categorical variables while maintaining high accuracy. CatBoost is known for its speed and performance on large datasets, making it ideal for AQI prediction tasks. By training on pollutant concentrations, meteorological features, and historical AQI data, CatBoost generates reliable forecasts, even in the presence of non-linear and multi-dimensional data relationships.
4) *LightGBM:* Light Gradient Boosting Machine (LightGBM) is another gradient boosting framework known for its high computational efficiency and scalability. LightGBM constructs decision trees in a leaf-wise manner, which reduces computation time and improves accuracy on large datasets. Its ability to handle missing data and outliers makes it well- suited for AQI prediction. By analysing historical air quality data and identifying patterns, LightGBM provides accurate predictions of AQI and facilitates better decision-making for air quality management.

Each regression model is trained and evaluated using key performance metrics, including R^2 scores and Root Mean Square Error (RMSE). These metrics help assess the predictive accuracy and robustness of the models, ensuring reliable AQI forecasting. By leveraging these advanced regression techniques, the model achieves a comprehensive understanding of pollutant behavior and its impact on air quality.

G. Evaluation Measures

To evaluate the performance of the regression models used for AQI prediction, two widely recognized metrics are employed: R^2 Score and Root Mean Square Error (RMSE). These metrics provide a comprehensive understanding of the model's accuracy and its ability to generalize to unseen data.

The R^2 Score measures the proportion of variance in the target variable (AQI) that is explained by the input features. It quantifies how well the regression model fits the data, with values ranging between 0 and 1. A higher R^2 score indicates that the model effectively captures the relationships between pollutants, meteorological factors, and AQI. In this project, R^2 is used to assess the accuracy of regression models such as Extra Trees Regressor, SVR, CatBoost, and LightGBM, ensuring that they align closely with the actual AQI values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

y_i: Actual observed value for the i[th] data point.
$\hat{y}_i$: Predicted value for the i[th] data point.
$\bar{y}$: Mean of the actual observed values.
n: Total number of data points.
$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Sum of squared errors (residual sum of squares).
$\sum_{i=1}^{n}(y_i - \bar{y})^2$: Total sum of squares, representing the variance in the observed data.

The Root Mean Square Error (RMSE) calculates the average deviation between the predicted AQI and the actual AQI. It emphasizes larger errors by squaring the differences, making it particularly useful for identifying models that consistently provide accurate predictions. A lower RMSE value signifies better model performance, as it indicates that the predictions are closer to the observed values. RMSE is critical for comparing models in this project, as it highlights the precision and reliability of the regression approaches.

Both metrics are integral to the evaluation process, offering insights into the predictive capabilities of the selected regression models. By analysing these measures, the most accurate and robust model can be chosen for AQI forecasting.

$$\text{RMSE} = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (3)$$

y_i: Actual observed value for the i[th] data point.
$\hat{y}_i$: Predicted value for the i[th] data point.
n: Total number of data points.
$(y_i - \hat{y}_i)^2$: Squared difference between actual and predicted values (squared residual).
$\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$: Sum of squared residuals.

IV. RESULTS

This section presents the findings from the multi-model regression approach developed for AQI prediction. It includes snapshots of model outputs, performance metrics, and a detailed evaluation of the prediction results.

A. Model Performance Comparison

The project evaluated four regression models: ExtraTrees Regressor, Support Vector Regression (SVR), CatBoost, and LightGBM, using metrics such as R-squared (R^2) and Root Mean Square Error (RMSE) across multiple cities.

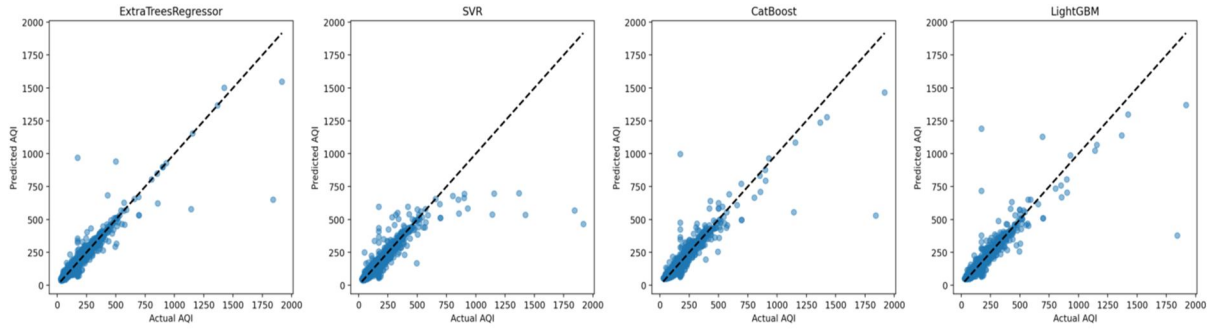


Figure 2. shows scatter plot representation of the AQI model.

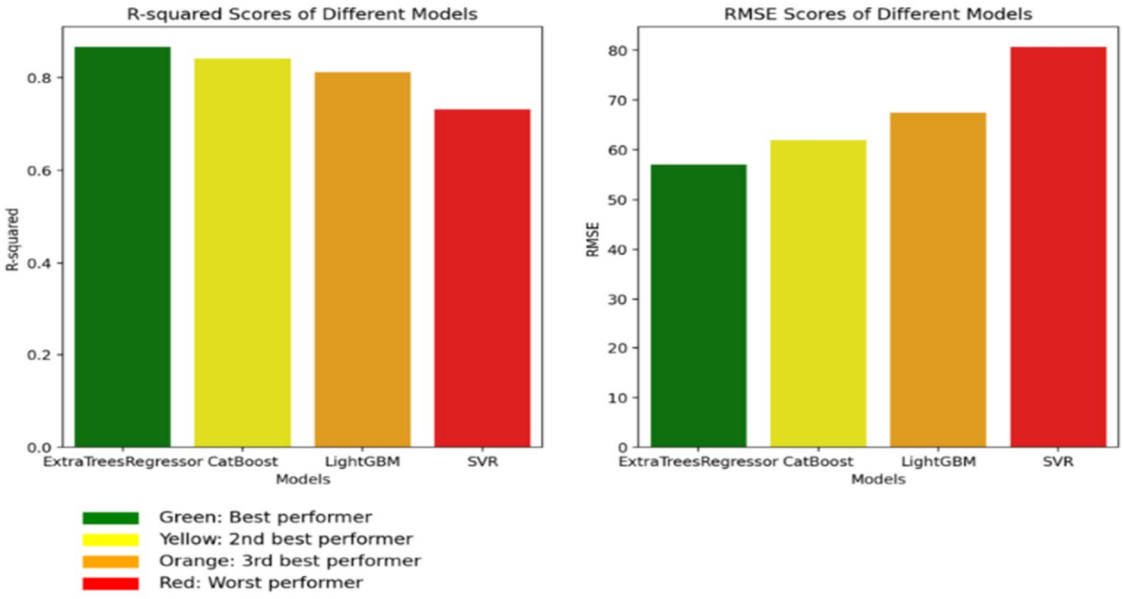


Figure 3. shows bar chart representation of the AQI model.

Model Performance Metrics:

	Model	R-squared	RMSE
0	ExtraTreesRegressor	0.8660	56.9836
1	SVR	0.7314	80.6680
2	CatBoost	0.8413	62.0080
3	LightGBM	0.8120	67.4816

Best Model:

Best model: **ExtraTreesRegressor** with R-squared value: **0.8660** and RMSE value: **56.9836**

Figure 4. shows the results of the AQI model.

This image presents a comparison of regression models based on R-squared and RMSE scores for the combined dataset of cities, identifying ExtraTrees Regressor as the best performer.

B. Visualizations Of Predictions

The performance of the regression models was assessed using scatterplots that compare predicted AQI values against actual AQI values. Instead of presenting multiple scatterplots for each model and city, Figure 5 is provided as a representative example. This scatterplot illustrates the predictive performance of the LightGBM model for a test city, with the predicted AQI values (y-axis) plotted against the actual AQI values (x-axis). The alignment of points near the diagonal dashed line reflects the model's accuracy, while deviations from this line highlight areas where the model's predictions can be improved.

This approach, utilizing scatterplots across all models, effectively conveyed the prediction accuracy and allowed for a straightforward evaluation of model performance across diverse data conditions.

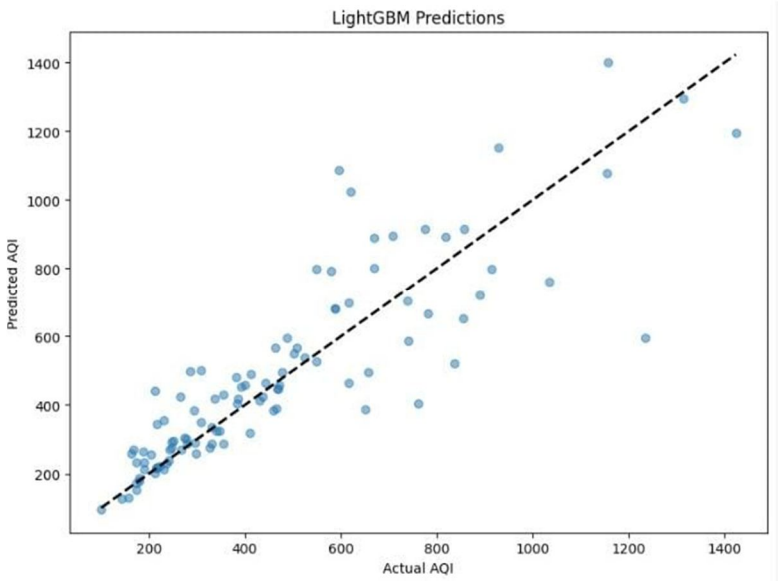


Figure 5. scatterplot for LightGBM predictions for a test city.

The performance of each model was further analysed using the obtained scatter plots for all the models and cities and came to a conclusion that represent as follows:

1) *ExtraTrees Regressor:* Scatter plots demonstrate that most points align closely with the diagonal, showing good prediction accuracy. Deviations at higher AQI values indicate slight overfitting.
2) *SVR:* Scatter plots reveal a broader scatter around the diagonal, especially for extreme AQI values. This suggests difficulty in capturing non-linear patterns.
3) *CatBoost:* Scatter plots indicate strong prediction performance, with most points clustered near the diagonal. Minor deviations suggest excellent generalizability.
4) *LightGBM:* Scatter plots exhibit the best alignment with the diagonal, showcasing the model's efficiency and robustness for AQI prediction.

C. CITY-WISE Analysis

The models were tested across cities, revealing interesting insights which is represented in Table 1:

1) *Ahmedabad:* Both CatBoost and LightGBM achieved high R^2 and low RMSE values, making them ideal for AQI prediction in this region.
2) *Delhi:* ExtraTrees Regressor showed superior performance in terms of R^2, but LightGBM performed better in RMSE, indicating its ability to handle outliers effectively.
3) *Bengaluru:* ExtraTrees achieved high R^2, while LightGBM excelled in minimizing prediction errors.
4) *Mumbai:* LightGBM emerged as the most balanced model, excelling in both R^2 and RMSE.

These findings suggest that the optimal model can vary based on regional and data-specific characteristics, underscoring the importance of multi-model evaluation.

Table 2. Result for best model according to different cities.

City	Best R^2 Model	Best RMSE Model
Ahmedabad	ExtraTrees Regressor	ExtraTrees Regressor
Delhi	ExtraTrees Regressor	ExtraTrees Regressor
Bengaluru	CatBoost	CatBoost
Mumbai	ExtraTrees Regressor	ExtraTrees Regressor

D. Insights From Feature Importance

Feature importance analysis revealed the following key predictors for AQI:

1) $PM_{2.5}$ and PM_{10}: Consistently the most significant features across all models.
2) Meteorological Variables: Factors like temperature, wind speed, and humidity moderately influenced predictions.
3) Gaseous Pollutants: NO_2 and SO_2 had varying levels of significance, contributing more in specific regions or during certain periods.

V. CONCLUSION

In this work, a robust and efficient Air Quality Index (AQI) prediction model has been developed using a multi-model regression approach. The methodology integrates data acquisition, preprocessing, and feature engineering to create a reliable foundation for accurate predictions. By leveraging advanced machine learning techniques such as SVR, ExtraTrees Regressor, CatBoost Regressor, and LightGBM Regressor, the work ensures that the model captures complex relationships between meteorological factors and air pollutants. Through rigorous evaluation based on metrics like accuracy, precision, and efficiency, ExtraTrees Regressor has been identified and proved to be the best-performing model. This model excels in capturing intricate patterns within the data while maintaining high computational efficiency.

The comprehensive pipeline ensures scalability and adaptability across various regions and environmental conditions. The work stands as a step forward in addressing air pollution challenges and promoting data-driven decision-making.

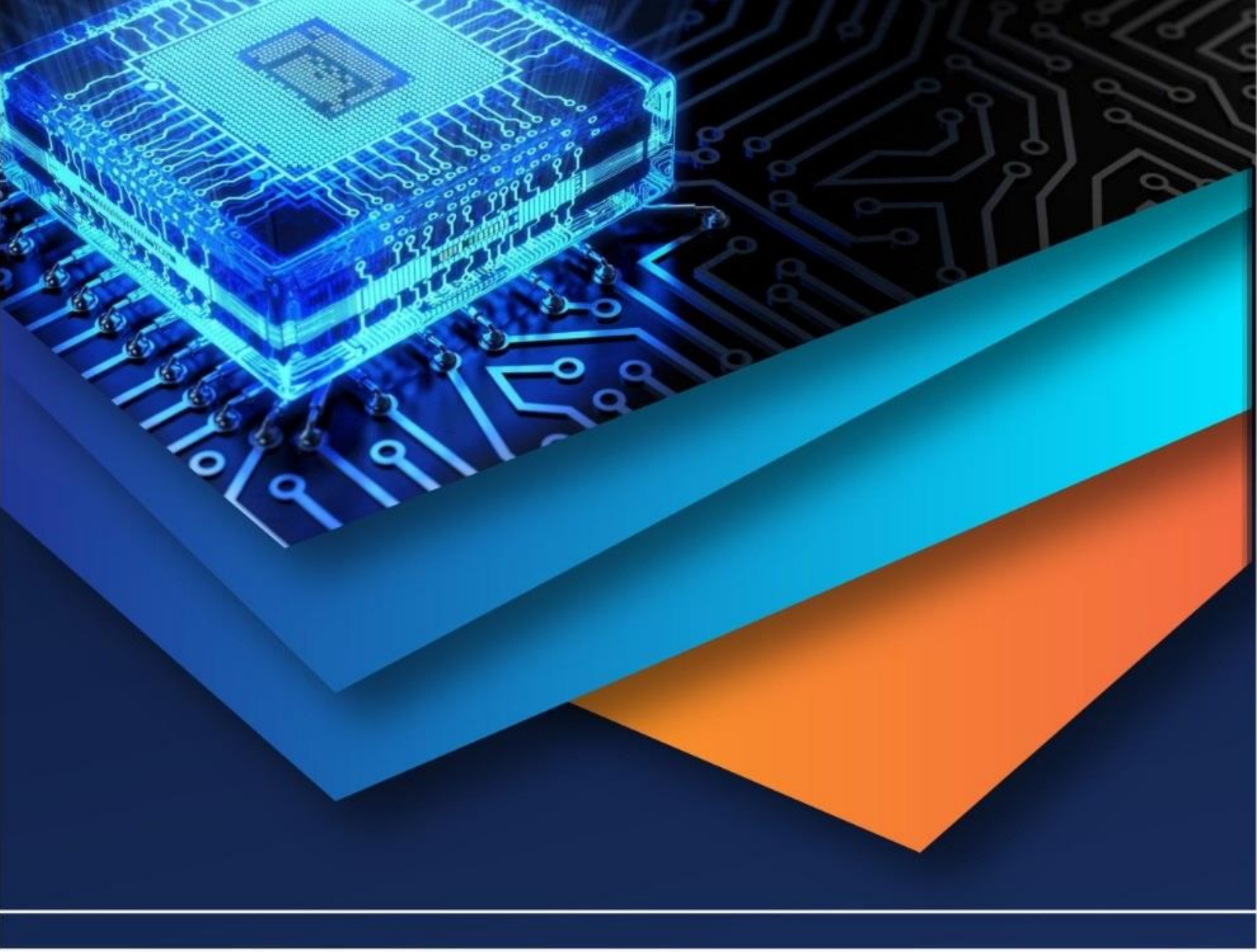
The final AQI prediction model, with the ExtraTrees Regressor at its core, can be utilized in real-world applications, such as urban planning, public health monitoring, and early warning systems for air pollution. Its ability to integrate real-time data and critical meteorological parameters makes it an essential tool for mitigating the adverse effects of poor air quality. Furthermore, the model's scalability ensures its effectiveness in diverse geographical areas, from densely populated cities to rural regions. By providing accurate and timely predictions, this system empowers policymakers, environmentalists, and individuals to take proactive measures.

The work highlights the importance of combining technological advancements with environmental awareness to create sustainable solutions. Future work could focus on incorporating additional factors like industrial emissions and socio-economic data to further improve prediction accuracy.

REFERENCES

[1] Ketu, Shwet. "Spatial air quality index and air pollutant concentration prediction using linear regression based recursive feature elimination with random forest regression (RFERF): a case study in India." Natural Hazards 114.2 (2022): 2109-2138.
[2] Mani, Geetha, and Joshi Kumar Viswanadhapalli. "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models." Journal of Engineering Research 10.2A (2022): 179-194.
[3] Kothandaraman, D., et al. "Intelligent forecasting of air quality and pollution prediction using machine learning." Adsorption Science & Technology 2022 (2022): 5086622.
[4] Suroshe, S., S. V. Dharpal, and N. W. Ingole. "Prediction of Air Quality Index Using Regression Models." GIS science journal 9.8 (2022): 576-591.
[5] 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.
[6] 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.
[7] Maltare, Nilesh N., and Safvan Vahora. "Air Quality Index prediction using machine learning for Ahmedabad city." Digital Chemical Engineering 7 (2023): 100093.
[8] Mahendra, H. N., et al. "Assessment and Prediction of Air Quality Level Using ARIMA Model: A Case Study of Surat City, Gujarat State, India." Nature Environment & Pollution Technology 22.1 (2023).
[9] Ravindiran, Gokulan, et al. "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam." Chemosphere 338 (2023): 139518.
[10] Wang, Yiyi, et al. "High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city." Environment International 172 (2023): 107752.
[11] Jumaah, Huda Jamal, et al. "Study of air contamination in Iraq using remotely sensed Data and GIS." Geocarto International 38.1 (2023): 2178518.
[12] Kazi, Zoltan, Snezana Filip, and Ljubica Kazi. "Predicting PM2. 5, PM10, SO2, NO2, NO and CO air pollutant values with linear regression in R language." Applied Sciences 13.6 (2023): 3617.
[13] Tang, Hao, et al. "A New Hybrid Forecasting Model Based on Dual Series Decomposition with Long-Term Short-Term Memory." International Journal of Intelligent Systems 2023.1 (2023): 9407104.
[14] Liang, Lu, et al. "Integrating low-cost sensor monitoring, satellite mapping, and geospatial artificial intelligence for intra-urban air pollution predictions." Environmental Pollution 331 (2023): 121832
[15] Zhang, Zhen, et al. "A systematic survey of air quality prediction based on deep learning." Alexandria Engineering Journal 93 (2024): 128-141.

[16] 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.
[17] Dalal, Surjeet, et al. "Optimising air quality prediction in smart cities with hybrid particle swarm optimization-long-short term memory-recurrent neural network model." IET Smart Cities (2024).
[18] Sun, Mingyue, Congjun Rao, and Zhuo Hu. "Air quality prediction using a novel three-stage model based on time series decomposition." Environment, Development and Sustainability (2024): 1-26.
[19] Suthar, Gourav, et al. "Predicting land surface temperature and examining its relationship with air pollution and urban parameters in Bengaluru: A machine learning approach." Urban Climate 53 (2024): 101830.
[20] Matthaios, Vasileios N., et al. "Predicting real-time within-vehicle air pollution exposure with mass-balance and machine learning approaches using on-road and air quality data." Atmospheric Environment 318 (2024): 120233.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY