Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 7s (2024)

Prediction of Air Quality Index and Air Quality Levels in Guwahati-India Using Machine Learning: A Comparative Study

[1]Joshua Remlalliana, [2*]Parthiban S

[1, 2*] Department of Mathematics and Statistics, School of Applied Sciences and Humanities, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur District, Andhra Pradesh, India, Pin: 522 213.

E-mail: [1]zosuaremlalliana@gmail.com; [2*]selvam.parthiban1979@gmail.com

Article History:

Received: *01-06-2024*

Revised: *03-07-2024*

Accepted: *29-07-2024*

Abstract:

This study provides a deep insight into the factors contributing to air pollution in Guwahati, India. It suggests measures for policymakers and urban planners to develop air quality management plans to control air pollution, not only for the benefit of human health but also for all living beings and the environment. The deteriorating air quality in urban areas, particularly in rapid growing city possesses notable health risks and environmental challenges. The main reason for examining the AQI (Air Quality Index) is the profound effects on health and environmental well-being. This research has analysed the evaluation and prediction of air quality based on the dataset obtained from Kaggle for the period of 2015-2020, which includes data on ten pollutants: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3 and Benzene. Three models from ML (Machine Learning), viz. DT (Decision Tree), RF (Random Forest), KNN (K-Nearest Neighbors) have been used for prediction and forecasting the AQI and AQL (Air Quality Index Levels). Finally, it has been observed that the RF Classification showed the highest accuracy in forecasting the AQL and factors such as PM10, PM2.5 and NH3 have been identified as the primary factors in determining AQI rating in Guwahati.

Keywords: Air Quality Index, Machine learning, Classification, Prediction, Forecasting.

AMS Classification: 62G08, 62J12, 62P12, 62M10, 91B76

1. Introduction

Clean air is crucial for sustaining life, protecting human health and preserving environmental integrity [3]. Air pollution can result in some serious diseases to human including bronchitis, pneumonia and various respiratory ailments [29]. Several critical factors influence the released of air contaminants in urban regions with respect to particulate matter, different gaseous emissions and sizes and several forms of meteorological conditions. This indicates that organizations, both governmental and non-governmental, across the world are crucial in developing the AQI [6]. Notably, $PM_{2.5}$ (Particulate Matter 2.5), PM_{10} (Particulate Matter 10), NO_2 (Carbon Dioxide), NO_x (Nitrogen Oxides), CO (Carbon Monoxide), SO_x (Sulphur Dioxides), O_3 (Ozone) and NH_3 (Ammonia) are stated to be responsible for the AQI indices. AQI levels above the stipulated limits can have adverse effects on the environment through global warming, occurrence of acid rain, smog, aerosols, visibility limitation and climatic changes [5]. The air pollution is increasing in urban areas due to high population density, rapid industrialization [32], the emission of industrial pollutants and vehicles [4]. Therefore, understanding and managing air quality are crucial for the Government and policymakers to develop effective management strategies.

AQI is a communication tool, which is used to convey the status of air in terms of conditions that are easily understandable to people [27]. It aggregates the information on several pollutants and for each pollutant, the data on concentration is converted to a number which is the index value, assigned a class and then to a corresponding colour. According to WHO [39], the governments should carefully consider their unique local conditions and the specific characteristics of each area when setting policy targets, particularly regarding the AQI. Central Pollution Control Board Standard is utilized to calculate AQI or Environment Pollution Index in India [21]. This index provides insights into the environmental condition regarding air quality, helping the general public understand the cleanliness or pollution level of the air. It provides valuable evaluation of air pollution for general public and helps in recognizing effective strategies or devices to lower the levels of major pollutants. The AQI represents the combined effect of all pollutants, giving a comprehensive overview of the overall air quality status. The break points for AQI scale (0-500) and categories of health statements are given in Table 1 [10], [37].

Table 1: Health guidelines based on AQI categories.

AQI Levels	Range	Correlated Health Outcomes
Good	**0-50**	Negligible influence.
Satisfactory	**51-100**	Mild irritation for individuals with a higher risk.
Moderate	**101-200**	Shortness of breath among children, elderly adults and people with comorbidities that affect the lungs or heart.
Poor	**201-300**	Shortness of breath in people who stayed for long periods and inconvenience for people with cardiac problems.
Very poor	**301-400**	Persistent impact leading to respiratory disease, especially for individuals with lung and heart ailments.
Severe	**401-500**	Impact on respiratory problems in healthy individuals and serious health consequences for those with lung or heart disease, even with light physical activity.

Guwahati, the leading city and capital of Assam, is the most extensive metropolitan region in the North-eastern Region of India [19]. It has experienced rapid urbanization and growth over the past few decades and this growth, driven by industrialization and a significant increase in vehicular traffic (87%) has severely impacted the city's air quality. As the gateway to nearly all north eastern states, Guwahati has seen a surge in fossil fuel-powered vehicles, contributing to rising pollution levels [13]. Recent studies have revealed a potential health threat in Guwahati in regard to air pollution with Hazard Quotient (HQ) estimated according to the norms of NAAQS and WHO which showed a significant likelihood of adverse health effects [15]. High levels of black carbon emissions, resulting from rapid urbanization and inadequate environmental regulations, have raised alarm bells in Guwahati [7]. Despite these challenges, there has been limited research focused on predictive modelling and health risk categorization of AQI in the city, making it a crucial area for further investigation.

In this present work, ML algorithms like RF (Random Forest), KNN (K-Nearest Neighbors) and DT (Decision Trees) have been performed for both regression and classification to predict AQI values and categorize health risk levels in Guwahati. Due to the skewed nature of the data, transformations are applied to upgrade the performance of this model and the results are scrutinized using performance indicators such as R^2 (Coefficient of Determination), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) for regression tasks. For classification, metrics such as Accuracy, F1 Score, Mathew's Correlation Coefficients (MCC) and AUROC (Area Under the Receiver Operating Characteristics) are implemented. The intent behind this research is to formulate a robust prediction mechanism for AQI and AQL and identify the main factors that significantly influence high AQI values in Guwahati so that the findings will be useful for the government and policymakers to take a preventive measure to reduce poor air quality. Thus, this paper aims to increase the accuracy of AQI predictions and strengthen air quality management in the region by identifying the critical polluting factors and applying innovative predictive models based on ML algorithms.

2. Literature Review

Gupta et al. (2023) used regression models [18] viz. SVR (Support Vector Regression), RFR (Random Forest Regression) and CR (CatBoost Regression) to forecast AQI in Indian cities such as New Delhi, Bangalore, Kolkata and Hyderabad. For data balancing, SMOTE (Synthetic Minority Oversampling Technique) algorithm was employed, which led to enhance model accuracy. The result shows that RFR provides the lowest values of RMSE and highest accuracy as compared to other techniques while CR provides the highest accuracy in some cities. Overall, the study demonstrated that RFR and CR models, especially when combined with dataset balancing techniques SMOTE, significantly improved the accuracy of AQI predictions in various Indian cities. The use of SMOTE in this study for dataset balancing might have its own limitations and assumptions that could affect the overall accuracy and reliability of the predictions and this paper lacks an in-depth discussion on the potential environment or health implications of inaccurate AQI predictions, which could be crucial for understanding the real-world impact of the study.

Dutta & Jinsart (2021a) compared different models [13], namely MLR (Multiple Linear Regression), ANN (Artificial Neural Network) and CART (Classification and Regression Trees) were used to predict PM_{10} concentrations at Guwahati city. Linear interpolation was applied for missing data and the result showed that ANN (MLP) was the best predictive performance compared to MLR and CART models for forecasting PM_{10} concentrations. This suggested that the non-linear algorithms like ANN can provide better predictive insights for air quality forecasting, aiding policymakers in decision-making processes. The research did not delve into the potential interactions between different pollutants like CO, NO_2, SO_2 and their combined effects on PM_{10} concentrations, which could provide a more comprehensive understanding of air quality dynamics.

K. Kumar & Pande (2023) highlighted the crucial need for monitoring and predicting air quality, attributing to the harmful impacts of industrial, transportation and household operations on the environment [29]. This research analysed atmospheric quality data collected over six years from 23 urban centres in Indian using ML techniques for prediction. Data pre-processing, feature selection, exploratory data analysis and resampling techniques were employed to address data imbalance. ML

models had been utilized, including Gaussian Naive Bayes, SVR and XGBoost for prediction with XGBoost model which outperformed other models. The paper acknowledged data imbalance problem and addressed it with resampling techniques, but the effectiveness of this approach in all scenarios was not extensively discussed.

Medhi & Gogoi (2021) analysed $PM_{2.5}$ concentrations in Guwahati city [36] before and after the lockdown using a visualization algorithm namely like time plots, heat maps, line charts and pie charts were employed to represent and analyse the data effectively. The result showed that the levels of $PM_{2.5}$ decreased during the lockdown period, indicating an improvement in air quality compared to previous years. This study did not delve into the specific sources of $PM_{2.5}$ pollution in Guwahati city, which could be a focus for future research. Zhou et al. (2018) introduced an integrated framework for characterizing and predicting AQI, considering individual pollutant concentrations for accurate AQI estimation [52]. NARX (Non-linear Autoregressive Neural Network with Exogenous Input) was employed and the outcome revealed the improvements in RMSE and MAPE values compared to alternative modelling techniques. This study demonstrated the effectiveness of NARX model in predicting AQI based on pollutant and meteorology data, showcasing improved performance over existing methods.

Ameer et al. (2019) used advanced regression techniques for predicting air pollution [4] and compared their performance based on MAE and RMSE. Experiments were performed using Apache Spark for pollution estimation with multiple datasets, evaluating processing time and error rates. The evaluation showed that Random Forest regression technique achieved the lowest error rates compared to other methods for most datasets. S. Zhu & X. Lian (2017) applied mixed models, EMD-SVR-Hybrid and EMD-IMFs-Hybrid [53], which utilized EMD for data cleaning and LS-SVR for forecasting the sum of IMFs, followed by S-ARIMA for residual sequence forecasting. The results demonstrated the superiority of EMD-SVR-Hybrid and EMD-IMFs-Hybrid models over SVR, ARIMA, GRNN and other methods in forecasting AQI data. In this study, the effectiveness of hybrid models might vary according to the characteristics of air quality data in different locations and computational complexity or scalability of the proposed hybrid models were not resolved.

Saikiran et al. (2021) focused on supervised ML algorithms [44] such as Linear Regression, SVR and RF to predict AQI with Random Forest showing higher accuracy due to lower over fitting. Liang et al. (2020) focused on predicting air quality using machine learning [30] like AdaBoost, ANN, random forest, stacking ensemble and SVM models based on 11 years dataset from TEPA (Taiwan Environmental Protection Agency). The analysis of the result depicted that Stacking Ensemble outperformed all other models in terms of R^2 and RMSE, while AdaBoost had the highest score in MAE, RF and ANN surpassed SVM for R^2 score. The limited longitudinal data might affect the model capacity to seasonal and other factors, potentially leading to the performance decay with increased time steps in certain regions which could affect the long-term forecasting accuracy. Mahalingam et al. (2019) employed SVM and ANN [33] to predict Delhi AQI. Kernel methods were used for identifying patterns and performing classification tasks. This study relied on historical data, which might not account for real-time fluctuations in air quality and the model performance could be influenced by accessibility and quality of the data.

3. Materials And Methods

3.1. ML algorithms for predicting AQI and AQL

Predicting AQI and AQL of Guwahati City, ML methods such as DT, RF and KNN are utilised. Fig. 1 shows the flowchart of predicting AQI and AQL.

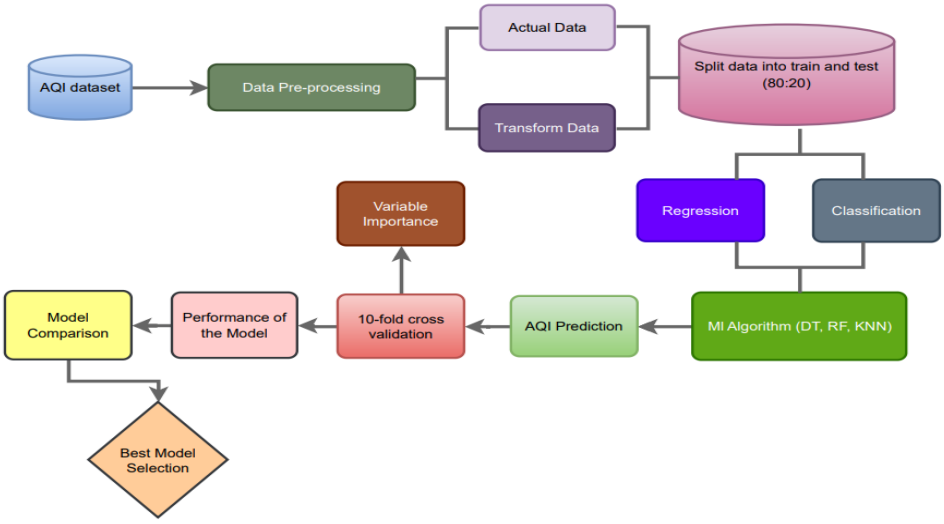


Fig. 1: Workflow for predicting AQI and AQL.

3.1.1. Decision Tree

DT is a statistical and ML methods applicable to both regression and classification problems [31], [50]. In regression, the Decision Tree algorithm focuses on giving a continuous AQI value in the output by minimizing the MSE at each split. DT model partitions the data at node *t* into two child nodes in a manner that minimizes RSS which helps in reducing the variance for better prediction [22]. The equation for RSS is formulated as:

$$RSS = \sum_{i \epsilon t} (u_i - \hat{u}_t)^2$$

where u_i represents AQI of the ith sample and $\hat{u}_t$ is the mean AQI of all samples in node *t*.

For classification, the DT algorithm classifies the AQI into discrete levels [31 by selecting the splits that minimize classification impurity measures such as Gini Impurity or Entropy [23]. The formula for Gini Impurity is defined by:

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i denotes the probability of data points being assigned to category *i*.

3.1.2. Random Forest

RF employs ensemble learning approach, systematically constructing multiple DT during the training process. For regression problems, it provides average of the prediction or majority vote for classification problems in an attempt to increase predictive accuracy and reduction of over-fitting [9],

[46]. RF algorithm estimates the continuous AQI value by averaging multiple decision trees built on different bootstrap samples using a random selection at each node [11], [45]. The equation for RF regression is given by:

$$\hat{u} = \frac{1}{T}\sum_{i=1}^{T} u_i$$

where $\hat{u}$ is the predicted value of AQI, T signifies the quantity of decision trees in RF and u_i represents the predicted value from the i^{th} tree.

For classification, the RF algorithm classifies the AQI levels by taking the majority vote from the individual decision trees. Each tree represents a class prediction and the final classification is the one that occurs most often across all the trees [17], [38]. RF classification can be expressed as:

$$\hat{c} = majority\ vote(c_1, c_2, \ldots., c_T)$$

where $\hat{c}$ is the final class prediction of the T^{th} decision tree.

3.1.3. K-Nearest Neighbors

KNN constitutes a supervised ML technique, with the ability to tackle classification as well as regression tasks [20], [51]. This method is often characterized as a lazy learning technique, as it requires little time for learning [26], [48]. KNN estimates the unknown value of a new data point by considering feature similarity, assigning a value from the training dataset proportional to its degree of similarity. The distance between the affected input and its neighbouring samples are quantified by a distance measure called Euclidean distance. This involves transforming the input dataset into a singular distance metric according to the computed similarity measures. This input is then classified or predicted based on the mean or mode of the labels of KNN with respect to classification or regression problem [12], [16]. The formula for measuring the Euclidean distance d seperating two points v and u is given by:

$$d(v,u) = \sqrt{\sum_{i=1}^{N}(v_i - u_i)^2}$$

where v_i represents the i^{th} feature of the input vector and u_i is the i^{th} feature of the training sample.

3.2. Model Evaluation

For the purpose of checking the accuracy and comparing the models, equations like RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (Coefficient of Determination), Accuracy, F1-score and MCC (Matthews Correlation Coefficient) were utilized with 10-fold cross validation for regression and classification model evaluation. Furthermore, as the classification metrics of the proposed models were derived on parts of the confusion matrix, ROCC (Receiver Operating Characteristic Curve) with AUROCC (Area Under ROC Curve) for each class using one-vs-rest method were assessed for the effectiveness of the model.

3.2.1. Root Mean Square Error

RMSE is a measure for assessing the accuracy of predictive models and it calculates the mean squared deviation of errors between the observed and forecasted values [1]. The formula that defines RMSE can be expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \hat{u}_i)^2}$$

where *n* represents overall sample size, u_i denotes the observed value for the i^{th} sample and $\hat{u}_i$ indicates the predicted value for the i^{th} sample.

3.2.2. Mean Absolute Error

MAE measures the average amount of deviation between the observed and predicted values [4]. It is calculated by summing the absolute variances between the actual and estimated data points, subsequently divided by the total sample size. The mathematical expression of MAE is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|u_i - \hat{u}_i|$$

where u_i indicates the actual value, $\hat{u}_i$ denotes the estimated value and *n* represents the overall number of data points.

3.2.3. Coefficient of Determination

R^2 expresses a degree in which variations of dependent variable are justified by the independent variables [18]. It provides a measure of how precisely the model predictions align with the real data by giving a value ranging from 0 to 1. It can be mathematically expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(u_i - \hat{u}_i)^2}{\sum_{i=1}^{n}(u_i - \bar{u}_i)^2}$$

where u_i signifies the actual value, $\bar{u}_i$ denotes the mean of the actual values and $\hat{u}_i$ indicates the predicted values.

3.2.4. Accuracy and F1 Score

Accuracy serves as a performance metric employed for evaluating the effectiveness behind a classification algorithm. This metric represents the ratio between correctly classified instances and all predictions made. Accuracy is calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The F1 Score is an evaluation metric constructed by blending the ideas of precision and recall into a singular value, calculated as the harmonic mean of these two measures with equal significance. It can be expressed as:

$$F1\ Score = 2\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$

- TP (True Positive): Instances that are correctly classified as positive.
- TN (True Negative): Instances that are correctly classified as negative.
- FP (False Positive): Instances that are incorrectly classified as positive.
- FN (False Negative): Instances that are incorrectly classified as negative.

3.2.5. Mathew's Correlation Coefficient

MCC is a metric that uses confusion matrix to determine the correlation connecting the actual and predicted class memberships of labels [8]. Like other correlation coefficients, MCC ranges from [-1, 1], where the lower end of the scale indicates poor performance, 0 indicates no correlation and 1 indicates an excellent feature in classifying the outcomes. The equation to calculate MCC is given by:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}}$$

3.2.6. Receiver Operating Characteristic Curve

The ROCC is a visual representation that plots the proportion of instances correctly classified as positive versus the proportion incorrectly classified as positive for multiple decision cut-off points. TPR is also referred to as sensitivity or recall while FPR is equal to 1- specificity. ROCC shows that while increasing TPR then FPR also increases conversely if TPR decreases then FPR also decreases. The AUROCC (Area Under ROC Curve) measure the area underneath the ROCC and gives a scalar value. An AUROCC close to 1 indicates good classification while values close to 0 suggest poor classification. As TPR or recall has already been given, the formula for FPR is expressed as:

$$FPR = \frac{FP}{FP + TN}$$

3.3. Data description and pre-processing

The current study utilizes Air Quality Data in India within the years 2015-2020 retrieved from Kaggle [2].

This dataset comprises hourly and daily AQI data collected at multiple monitoring sites distributed throughout different Indian municipalities, namely Ahmedabad, Aizawl, Amaravati, Amritsar, Bangalore, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam. The daily data samples consist of 29,532 rows and 16 columns while the hourly dataset comprises 707,876 rows and 16 columns. The columns in these datasets are City, Datetime, $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, Toluene, Xylene, AQI, and AQI_Bucket. The AQI_Bucket has six levels such as good, satisfactory, moderate, poor, very poor, and severe. The data has been filtered to isolate Guwahati due to its status as the most polluted city in North-eastern Region of India, emphasizing the importance of analysing

its pollution levels. In this study the hourly dataset is selected for detailed analysis, resulting in a subset containing 12,003 rows and 16 columns. Cleaning of the data involved removing the parameters Toluene and Xylene using Microsoft Excel, as these columns contained entirely empty values. Subsequent pre-processing steps are conducted using R programming software to identify and eliminate NA and missing values. After this cleaning process, the dataset for Guwahati is refined to include 11,413 rows and 14 columns. Further analysis is done using JASP software.

3.4. Data exploration

This section of the present study focuses on EDA (Exploratory Data Analysis) that is important for identification of the hidden patterns in the dataset before applying ML algorithm. Descriptive statistics for the air pollutants and AQI are computed to give an overview of the data. The central tendency, dispersion and range of the data are presented in Table 2. Fig. 2 presents the pie chart of AQI levels in the dataset. It is also observed that all the parameters and the AQI are positively skewed as shown in Fig. 3, thus indicating that the data was not normally distributed. Additionally, to understand the relationship between the various air contaminants and how they affect the AQI, Spearman's correlation heat map is created and shown in Fig. 3. This heat map ranges from -1 to +1, with colour intensity representing the degree of correlation. From this correlation heat map, it is observed that the correlation coefficient between AQI and PM_{10} is 0.772 while AQI and $PM_{2.5}$ is 0.752. Similarly, 0.692 for NH_3, 0.534 for CO, 0.532 for NO_x and 0.518 for Benzene. So, correlation heat map indicates that PM_{10} and $PM_{2.5}$ followed by NH_3, CO, NO_x and Benzene emerge as the primary drivers in estimating the AQI values. Among the air pollutants, $PM_{2.5}$ has a strong relationship with PM_{10} as both falls under the category of particulate matter. The presence of strong association is also found between NO and NO_x.

Table 2: Summary statistics of air quality indicators.

Attributes	Count	Mean	Std. Dev	Min. value	IQR1	IQR2	IQR3	Max. value
$PM_{2.5}$	11413	61.364	70.823	0.06	19	38.08	80.31	923.08
PM_{10}	11413	112.882	134.467	1.11	33.73	68.23	141.42	1000
NO	11413	19.888	31.743	0.21	1.62	4.57	21.66	185.27
NO_2	11413	13.664	10.925	1.35	6.83	10.18	16.32	107.04
NO_x	11413	44.091	56.678	0.81	9.38	19.09	50.59	319.65
NH_3	11413	10.794	7.658	0.11	4.45	8.79	16.22	47.98
CO	11413	0.734	0.617	0	0.37	0.51	0.88	6.6
SO_2	11413	14.577	4.697	3.91	11.82	13.97	15.95	140.84
O_3	11413	25.127	17.763	0.28	17.32	18.38	27.79	163.17
Benzene	11413	2.386	18.721	0	0	0.19	0.92	491.51
AQI	11413	140.026	115.633	18	52	98	213	1109

Fig. 2: Pie-chart of AQI levels in the datasets.

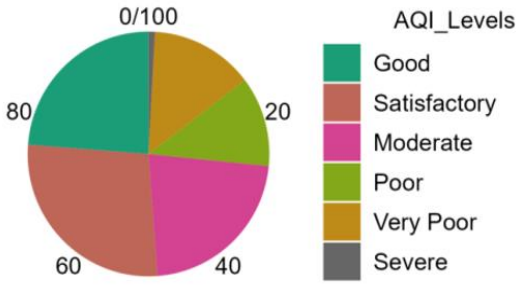
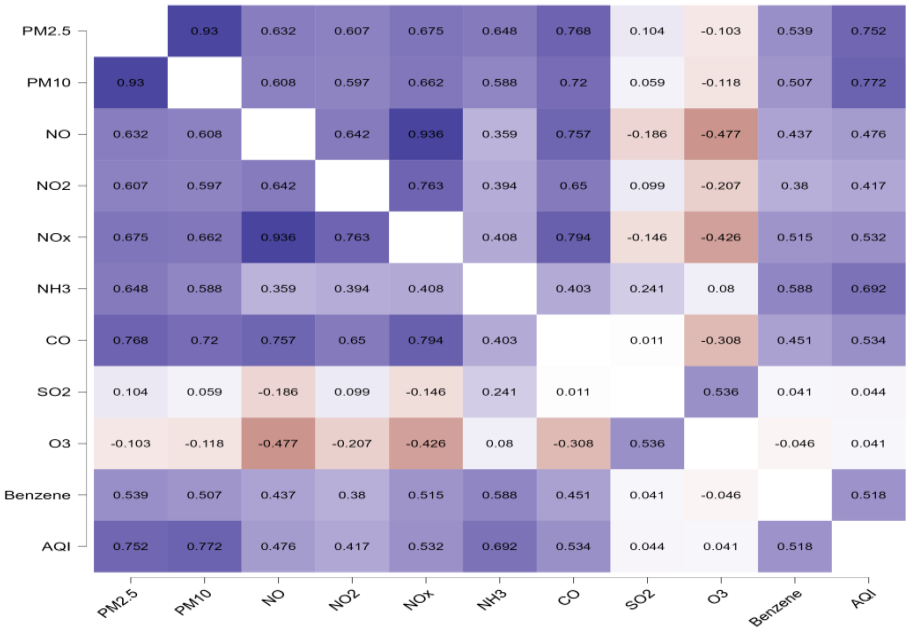


Fig. 3: Heat map of correlation matrix.



3.5. Transformation of data

To achieve better results with ML methods, working with normally distributed data is crucial, especially when dealing with positively skewed datasets. In this study, the distributions of different air pollutants and AQI have been transformed using three distinct methods: a square-root transformation, a log-normal transformation and a Box-Cox transformation. Among these methods, the transformation that gives values closest to a normal distribution has been selected based on skewness and kurtosis metrics. The acceptable range of skewness is usually from -3 and +3, while the acceptable range for kurtosis is between -10 and +10 [43]. These ranges indicate that the data distribution is almost normal. Table 3 illustrates asymmetry (skewness) and peakedness (kurtosis) measures of key atmospheric contaminants with and without applying transformations. Fig. 4 and Fig. 5 illustrates the transformation showing a clear enhancement in the distribution of the data.

Table 3: Comparison of skewness and kurtosis with and without transformations.

Attributes	Without Transformation		With Transformation	
	Skewness	Kurtosis	Skewness	Kurtosis
$PM_{2.5}$	3.983	28.258	-0.25	0.752
PM_{10}	3.429	16.221	-0.042	-0.289

NO	2.097	3.87	-0.188	-0.882
NO_2	2.471	8.308	0.45	-0.075
NO_x	1.931	3.146	0.027	-0.676
NH_3	0.825	0.247	-0.059	-0.839
CO	2.801	10.915	0.014	2.609
SO_2	5.869	81.651	-0.1	2.296
O_3	1.938	4.34	0.052	1.593
Benzene	14.644	257.663	0.008	-1.208
AQI	2.006	9.21	0.021	-1.106

Fig 4: Air pollutants and AQI distributions without transformation.

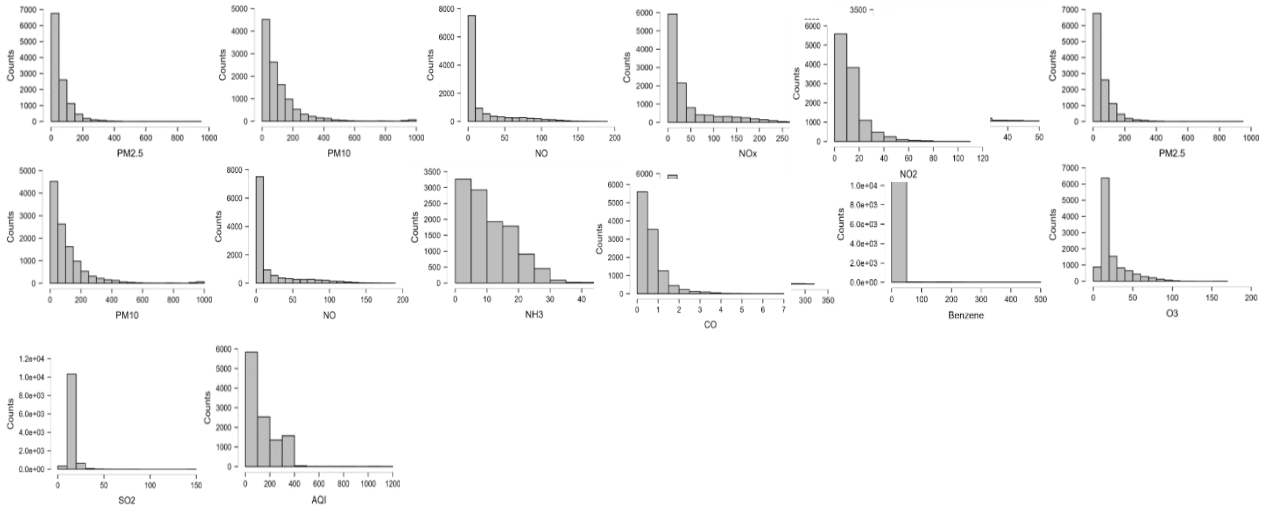
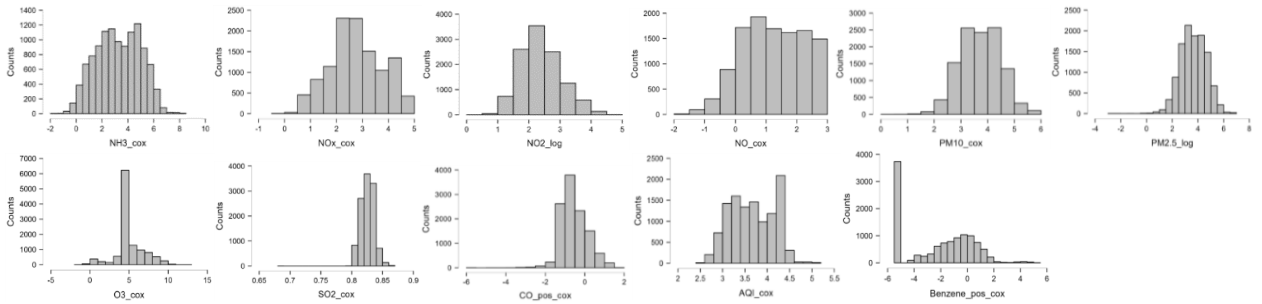


Fig 5: Air pollutants and AQI distributions with transformation.



3.6. Splitting of data

To ascertain the model efficiency, the entire set of data has been partition into training and testing datasets. For regression and classification analysis, the training and testing sets are split in a ratio of 80% to 20%. The initial 80% of the data has been utilised to develop and estimate the model parameters, while the remaining 20% has been used for testing the developed model and performance evaluation.

4. Results And Discussion

4.1. Summary of data

In this work, the air quality of Guwahati dataset during the year (2015-2020) has been analysed and cleaned before applying ML algorithms, data exploration has been done to explore the hidden pattern present in the dataset, descriptive statistics are shown in Table 2 and Fig. 2 represents the pie chart of AQL. It is observed that most of the pollutants have having large standard deviations, indicating high variability. The median values (50%) of all attributes are lower than the mean values, suggesting a right-skewed distribution as depicted in Fig. 4. The interquartile ranges are relatively narrow as compared to the overall range and the presence of extreme values in PM_{10} and $PM_{2.5}$ indicates significant fluctuations in pollution levels.

The highest recorded levels of air pollutants and AQI values in this research exceed NAAQ Standards of India [37]. Particulate matter PM_{10} and $PM_{2.5}$, in addition to atmospheric gases, pose a major public health issue where the concentration is higher than its standard limit set by the authorities. The effects of atmospheric pollutants on human health vary based on the specific pollutant, its environmental concentration and the time span of exposure. These concentrations are influenced not only by the quantity released from sources but also by the atmosphere's capacity for absorption or dispersion of these emissions [47], [49]. Different pollutants can vary in their levels of exposure; for instance, PM_{10} and $PM_{2.5}$ can significantly enter respiratory and cardiovascular systems of the human body. These pollutants are known to trigger asthma attacks, bronchitis, reduce lung function and increase the risk of heart attacks. Due to their small size, they can enter the bloodstream, exacerbating chronic diseases and raising the likelihood of mortality [49]. NH_3 (Ammonia) can cause irritation to the eyes, throat and the lungs and it results in cough, wheezing and breathlessness. It can lead to the aggravation of asthma and other respiratory diseases, as well as eutrophication of waters bodies [28]. CO (Carbon Monoxide) blocks oxygen to deliver to body tissues by binding with haemoglobin that affects cardiovascular and neurobehavioral systems. Long-term effects involve changes in irregular heartbeat, heart disease and neurological disorders [25]. NO_x (Nitrogen Oxides) particularly NO_2 (Nitrogen Dioxide) and NO (Nitric Oxide), contribute to asthma and reduced lung function. They also play a role in smog formation and generation of acid rain, leading to ecosystem acidification, biodiversity loss and degradation of water bodies, soil, and vegetation [25]. SO_2 (Sulfur Dioxide) irritates the respiratory system and can also stimulate other diseases such as Asthma and Bronchitis. It is also a source of acid rain that in turns affects vegetation, water sources, wildlife and infrastructure [34]. O_3 (Ozone) can be inhaled and it can affect the respiratory system wherein it can cause functional changes in pulmonary function, increases the airway inflammation and increase in the sensitivity to broncho-constrictors [47]. Benzene is also found associated with blood disorders such as leukaemia. It is a carcinogen that has an impact on the bone marrow that results into anaemia and immunosuppression [35].

The major air pollutants in Guwahati significantly harm both public health and the environment [14]. Tolerance to these pollutants may cause different respiratory disorders, cardiovascular ailments and other chronic diseases in people, particularly children, the elderly and patients with severe health complications. Measures to control and reduce air pollution in Guwahati is crucial for ensuring public health, protecting the environment and enhancing the city's development. This would

necessitate the use of policies, decreasing atmospheric emissions, public awareness campaigns and cooperation to the government, industries and the community. The present study aims at identifying the factors needed for AQI prediction for Guwahati and parameters to be incorporated for developing an accurate AQI prediction model to support air pollution management. These models can give current occurrence and future predictions of air quality, thereby assisting authorities in implementing strategies to reduce the impacts of air pollution. Using these factors for models that forecast AQI in Guwahati, the government and other authorities can decide on controlling measures like developing a traffic management plan, setting the standards of industrial emissions and encouraging the utilisation of clean energy resources. This will help them be in a position to respond adequately to air quality matters and to ensure public health in the true sense for people in the city is maintained.

4.2. Comparison for AQI models

Table 4 showed the comparison of ML models for predicting AQI between actual data and transformed data. The model evaluations utilized for this study were R^2, RMSE and MAE. Among these three models on actual data, KNNR scores the highest value of R^2 equal to 0.719, RMSE of 62.384 and MAE of 31.233 as compared to the other two models while DTR exhibited the lowest score in R^2 of 0.673 and the highest score in RMSE of 67.372 and MAE of 35.845. RFR lies between the two models displayed slightly lower performance than KNNR, scoring R^2 of 0.703, RMSE of 64.047 and MSE of 33.21. In the context of transformed data, KNNR continues to outperform RFR and DTR models with the highest score in R^2 value of 0.837, RMSE of 0.202 and MAE of 0.136. The RFR model closely reflects KNNR results, yielding R^2 equal to 0.812, RMSE of 0.217 and MAE of 0.151, while DTR remains the lowest performing model scoring R^2 of 0.757, RMSE of 0.247 and MAE of 0.174. As shown in Fig. 5, increases in R^2 values in the transformed data consistently improves the performance of the three models and also decreases in RMSE and MAE values as compared to actual data. This indicates that data transformation significantly amplifies the accuracy and robustness of the model output. KNNR outperformed the other two models with the highest score in R^2, least score in RMSE and MAE in both actual and transformed datasets. Figs. 7, 8 and 9 demonstrates the predictive performance of both actual and transformed data. Due to its best performance, feature importance for predicting AQI for KNNR model is shown in Fig. 6. From this feature importance figure, it has been determined that PM_{10} and $PM_{2.5}$ are the most significant factors for determining AQI. This indicates the need to enhance the regulation of particulate emissions to improve atmospheric conditions in the city. Some important gaseous pollutants, which help in enhancing the primary as well as secondary air pollutants are NH_3 , CO, NO_x , NO, SO_2, O_3, Benzene and NO_2. Most of the priority pollutants identified in this study are confirmed to affect human health and play a vital part in the deterioration of air quality [14]. With these contributions in mind, it becomes easier for the authorities to focus on methods to control emissions and ways of enhancing the health of people in Guwahati by foster better air quality.

Table 4: Comparison of ML regression models in actual and transform data.

Models	Actual Data			Transform Data		
	R^2	RMSE	MAE	R^2	RMSE	MAE

DTR	0.673	67.372	35.845	0.757	0.247	0.174
RFR	0.703	64.047	33.21	0.812	0.217	0.151
KNNR	0.719	62.384	31.233	0.837	0.202	0.136

Fig. 5: Comparison of R^2 for actual and transformed data.

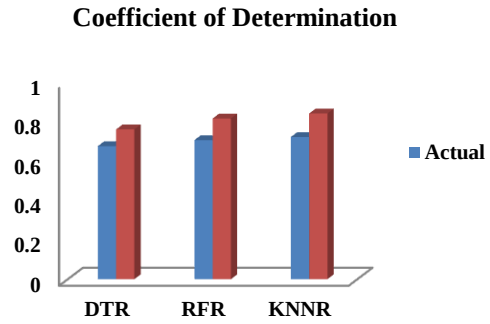


Fig 6: Feature Importance for forecasting AQI in KNNR.

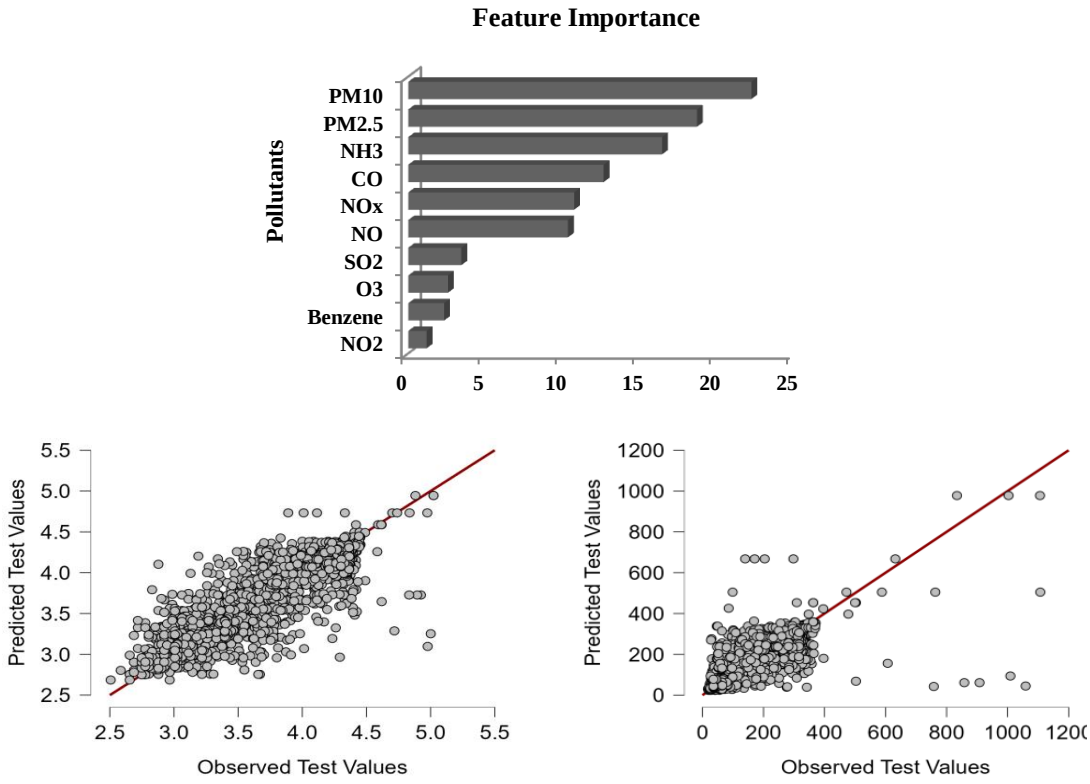


Fig. 7: Comparison of predictive performance plot in DTR model for actual and transform data.

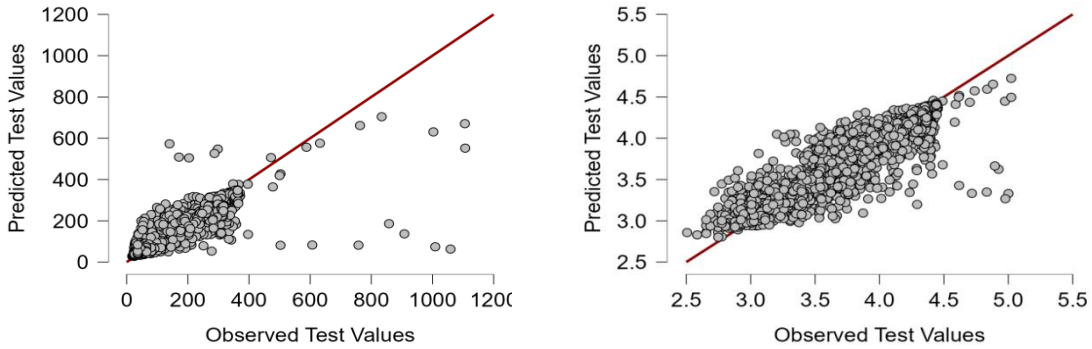


Fig. 8: Comparison of predictive performance plot in RFR model for actual and transform data.

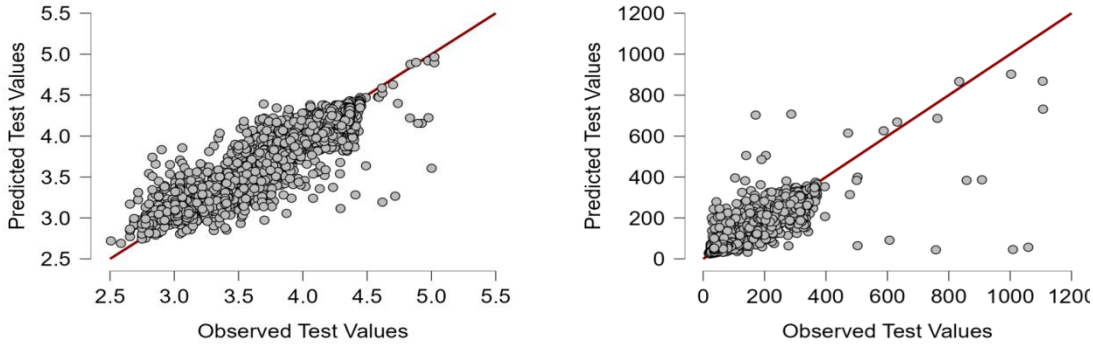


Fig. 9: Comparison of predictive performance plot in KNNR model for actual and transform data.

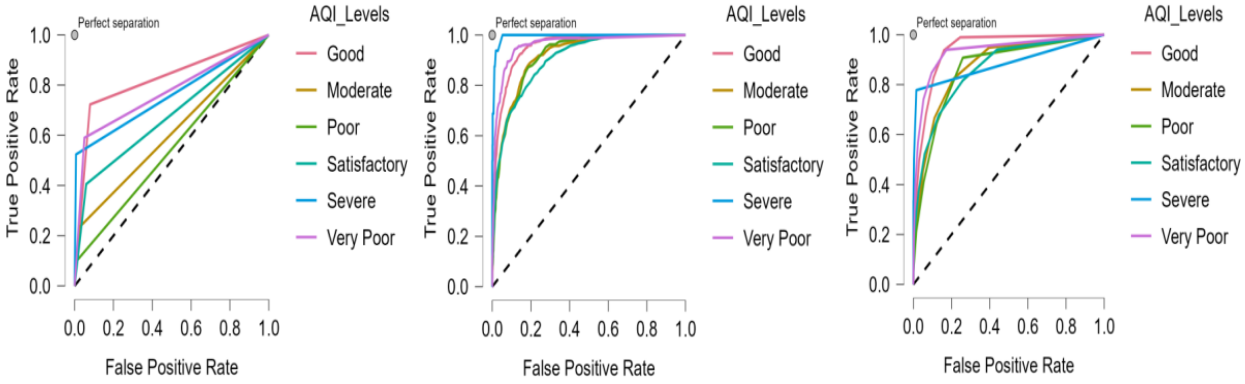
4.3. Comparison for AQL models

From Table 5, RFC demonstrates the highest accuracy of 90.8%, F1 Score of 0.718, MCC of 0.749 and the highest AUC of 0.937 indicates that RFC is doing a good job in classifying AQL. KNNC ranks as the second-best performing model, slightly behind RFC with a decent accuracy of 89.7%, F1 score of 0.689 and a slightly higher MCC of 0.754, suggesting that while its class prediction balance is marginally better than RFC and it falls short in precision and recall balance. Additionally, it has a commendable AUC of 0.893. DTC has the least performance among the three with an accuracy of 87.5%, F1 score of 0.606, indicating less reliable balance between precision and recall and the lowest MCC of 0.51, which suggests more misclassifications compared to the other models. AUC of 0.746, although lower, still indicates a fair distinction capacity between classes. Fig. 10, which plots the ROC curve for each model using the one-vs-rest method, demonstrates their respective average AUC values, further emphasizing RFC's effectiveness in class prediction in the evaluation metrics. Fig. 11 represents a comparison of the classification model based on their evaluation matrices. Fig. 12 shows the feature importance by mean decrease in accuracy plot of the best model for forecasting the AQL. From this plot, it has been found that PM_{10}, $PM_{2.5}$ and NH_3 are the key factors that contribute in classifying the AQI levels.

Table 5: Performance of ML classification models

Models	Model evaluation			
	ACC	F1	MCC	AUC
DTC	0.875	0.606	0.51	0.746
RFC	0.908	0.718	0.749	0.937
KNNC	0.897	0.689	0.754	0.893

Fig. 10: ROC Curves for (a) DTC, (b) RFC and (c) KNNC



(a) (b) (c)

Fig. 11: Comparison of ML classification models

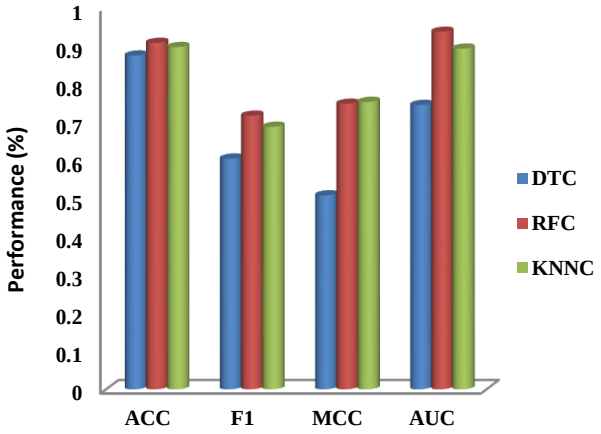
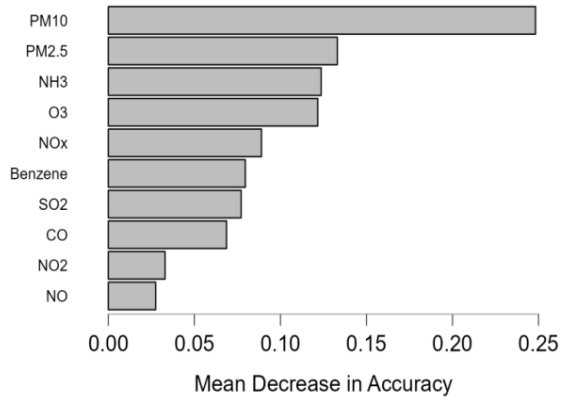


Fig. 12: Feature Importance for forecasting AQI levels in RFC



4.4. Implications

This research on predicting AQI and AQI levels in Guwahati using ML algorithms are profound for both public health and urban planning. Accurate prediction of AQI and identification of key contributing factors are important in safeguarding human health from the effects of pollution. After analysing the results from regression and classification methods, it has been observed that KNNR and RFC models are effective in predicting AQI values and classifying AQL. Identifying PM_{10}, $PM_{2.5}$ and NH_3, as the most pollutant factors, contributing to high AQI allows the government and policymakers to prioritize interventions on these pollutants. The prediction model for air pollution support preventive measures such as warnings on high pollution days, traffic control measures and use of public transport in order to reduce emissions while features importance analysis supports environmental policies and city planning so that these growths in Guwahati do not worsen air pollution. By providing a robust analytical framework, this research supports the formulation of data-driven plans for enhancing air quality, thereby promoting the well-being of the city's residents.

5. Conclusion And Future Work

This study has demonstrated a comparison of ML algorithms like RF, KNN and DT for predicting the AQI and AQL for regression and classification in Guwahati during the year 2015 to 2020. Among the regression models, KNNR outperformed RF and DT with R^2 equal to 0.812, RMSE of 0.217 and MAE of 0.151. For classification, RFC has the highest accuracy of 90.8%, F1 Score of 0.718, MCC of 0.749 and AUC of 0.937. The results of feature importance revealed that PM_{10}, $PM_{2.5}$ and NH_3 are the three most critical pollutants influencing AQI and AQI levels. The insights provided by this research can inform government policies and regulatory frameworks aimed at reducing pollution levels and protecting public health. Overall, this study reveals the capability of ML technique in improving air quality management and underscores importance of continuous monitoring and adaptation of predictive models to reflect changing environmental conditions. Thus, as for the future research, this study can be carried on to the higher level by applying other ML techniques such as SVM (Support Vector Machines), ANN (Artificial Neural Network) and GBM (Gradient Boosting Machines) to improve predictive capability of AQI and AQL. Moreover, there is a possibility to increase the predictive accuracy by utilising replication techniques including stacking, bagging and boosting. Further, the integration of meteorological data like wind speed,

humidity and temperature will enhance consideration of many more influences on the air quality and improve prediction model.

REFERENCES

[1] Ahmadi, K., Kalantar, B., Saeidi, V., Harandi, E. K. G., Janizadeh, S., & Ueda, N. (2020). Comparison of Machine Learning Methods for Mapping the Stand Characteristics of Temperate Forests Using Multi-Spectral Sentinel-2 Data. *Remote Sensing*, *12*(18), 3019. https://doi.org/10.3390/rs12183019

[2] *Air Quality Data in India (2015—2020)*. (n.d.). Retrieved May 24, 2024, from https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india/data

[3] Akinfolarin, O. M., Boisa, N., & Obunwo, C. C. (2017). Assessment of Particulate Matter-Based Air Quality Index in Port Harcourt, Nigeria. *Journal of Environmental Analytical Chemistry*, *04*(04). https://doi.org/10.4172/2380-2391.1000224

[4] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. *IEEE Access*, *7*, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

[5] Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., Stanaway, J. D., Beig, G., Joshi, T. K., Aggarwal, A. N., Sabde, Y., Sadhu, H., Frostad, J., Causey, K., Godwin, W., Shukla, D. K., Kumar, G. A., Varghese, C. M., Muraleedharan, P., … Dandona, L. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. *The Lancet Planetary Health*, *3*(1), e26–e39. https://doi.org/10.1016/S2542-5196(18)30261-4

[6] Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. *Science of The Total Environment*, *731*, 139052. https://doi.org/10.1016/j.scitotenv.2020.139052

[7] Barman, N., & Gokhale, S. (2019). Urban black carbon—Source apportionment, emissions and long-range transport over the Brahmaputra River Valley. *Science of The Total Environment*, *693*, 133577. https://doi.org/10.1016/j.scitotenv.2019.07.383

[8] Boughorbel, S., Jarray, F., & El-Anbari, M. (n.d.). *Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric*.

[9] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Central Pollution Control Board. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Central_Pollution_Control_Board&oldid=1225414990

[11] Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification In Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

[12] Danades, A., Pratama, D., Anggraini, D., & Anggriani, D. (2016). Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status. *2016 6th International Conference on System Engineering and Technology (ICSET)*, 137–141. https://doi.org/10.1109/ICSEngT.2016.7849638

[13] Dutta, A., & Jinsart, W. (2021a). Air Pollution in Indian Cities and Comparison of MLR, ANN and CART Models for Predicting PM10 Concentrations in Guwahati, India. *Asian Journal of Atmospheric Environment*, *15*(1), 68–93. https://doi.org/10.5572/ajae.2020.131

[14] Dutta, A., & Jinsart, W. (2021b). Assessing short-term effects of ambient air pollution on respiratory diseases in Guwahati, India with the application of the generalized additive model. *Human and Ecological Risk Assessment: An International Journal*, *27*(7), 1786–1807. https://doi.org/10.1080/10807039.2021.1908113

[15] Dutta, A., & Jinsart, W. (2021c). Risks to health from ambient particulate matter ($PM_{2.5}$) to the residents of Guwahati city, India: An analysis of prediction model. *Human and Ecological Risk Assessment: An International Journal*, *27*(4), 1094–1111. https://doi.org/10.1080/10807039.2020.1807902

[16] Ertuğrul, Ö. F., & Tağluk, M. E. (2017). A novel version of k nearest neighbor: Dependent nearest neighbor. *Applied Soft Computing*, *55*, 480–490. https://doi.org/10.1016/j.asoc.2017.02.020

[17] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., & Rueckert, D. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, *65*, 167–175. https://doi.org/10.1016/j.neuroimage.2012.09.065

[18] Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal of Environmental and Public Health*, *2023*, 4916267. https://doi.org/10.1155/2023/4916267

[19] Guwahati. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Guwahati&oldid=1223066769

[20] Hamed, Y., Ibrahim Alzahrani, A., Shafie, A., Mustaffa, Z., Che Ismail, M., & Kok Eng, K. (2020). Two steps hybrid calibration algorithm of support vector regression and K-nearest neighbors. *Alexandria Engineering Journal*, *59*(3), 1181–1190. https://doi.org/10.1016/j.aej.2020.01.033

[21] Jain, R. K., Kori, R., & Saxena, A. (2010). Ambient air quality status of Bhopal, Madhya Pradesh. *Studies on Pollution Mitigation. CPCB Pp*, *3944*.

[22] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021a). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

[23] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021b). Classification. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 129–195). Springer US. https://doi.org/10.1007/978-1-0716-1418-1_4

[24] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021c). Tree-Based Methods. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 327–365). Springer US. https://doi.org/10.1007/978-1-0716-1418-1_8

[25] Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, *151* (2), 362–367. https://doi.org/10.1016/j.envpol.2007.06.012

[26] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

[27] Kowalska, M., Ośródka, L., Klejnowski, K., Zejda, J. E., Krajny, E., & Wojtylak, M. (2009). Air quality index and its significance in environmental health risk communication. *Archives of Environmental Protection*, 13–21.

[28] Kumar, A., Patil, R. S., Dikshit, A. K., & Kumar, R. (2019). Assessment of Spatial Ambient Concentration of NH3 and its Health Impact for Mumbai City. *Asian Journal of Atmospheric Environment*, *13*(1), 11–19. https://doi.org/10.5572/ajae.2019.13.1.011

[29] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: A case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333–5348. https://doi.org/10.1007/s13762-022-04241-5

[30] Liang, Y.-C., Maimury, Y., Chen, A. H.-L., & Juarez, J. R. C. (2020). Machine Learning-Based Prediction of Air Quality. *Applied Sciences*, *10*(24), 9151. https://doi.org/10.3390/app10249151

[31] Loh, W. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

[32] Madan, T., Sagar, S., & Virmani, D. (2020). Air Quality Prediction using Machine Learning Algorithms –A Review. *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 140–145. https://doi.org/10.1109/ICACCCN51052.2020.9362912

[33] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). A Machine Learning Model for Air Quality Prediction for Smart Cities. *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 452–457. https://doi.org/10.1109/WiSPNET45539.2019.9032734

[34] Maynard, R. (2004). Key airborne pollutants—The impact on health. *Science of The Total Environment*, *334–335*, 9–13. https://doi.org/10.1016/j.scitotenv.2004.04.025

[35] McHale, C. M., Zhang, L., & Smith, M. T. (2012). Current understanding of the mechanism of benzene-induced leukemia in humans: Implications for risk assessment. *Carcinogenesis*, *33*(2), 240–252. https://doi.org/10.1093/carcin/bgr297

[36] Medhi, S., & Gogoi, M. (2021). Visualization and Analysis of COVID-19 Impact on PM2.5 Concentration in Guwahati city. *2021 International Conference on Computational Performance Evaluation (ComPE)*, 012–016. https://doi.org/10.1109/ComPE53109.2021.9752244

[37] *NAAQS_2019.pdf*. (n.d.). Retrieved June 1, 2024, from https://cpcb.nic.in/upload/NAAQS_2019.pdf

[38] Njoku, O. (2019). Decision Trees and Their Application for Classification and Regression Problems. *MSU Graduate Theses*. https://bearworks.missouristate.edu/theses/3406

[39] Organization, W. H. (2006). WHO. Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide. Global update 2005. *World Health Organization. Available from: Http://Www. Euro. Who. Int/__data/Assets/Pdf_file/0005/786*, *38*, E90038.

[40] Parthiban, S. and Gajivaradhan, P., A comparative study of statistical hypothesis test for 2^2 factorial experiments under fuzzy environments, Bulletin of Mathematics and Statistics Research, Vol. 4, Issue: 1, (January-March 2016), pp. 46-70.
[41] http://www.bomsr.com/4.1.16/46-70%20P.%20GAJIVARADHAN.pdf
[42] Parthiban, S. and Gajivaradhan, P., Statistical hypothesis test in three factor ANOVA model under fuzzy environments using trapezoidal fuzzy numbers, Bulletin of Mathematical Sciences and Applications, Vol. 14, (2016), pp. 23-42.
[43] https://www.scipress.com/BMSA.14.23
[44] Parthiban, S. and Gopinathan, P., Statistical Hypothesis on Industrial Applications through Ranks from COG of TrFNs, International Journal of Recent Technology and Engineering (IJRTE), Vol. 6, Issue: 1S4, (June 2019), pp. 1116-1118.
[45] https://www.ijrte.org/download/volume-8-issue-1s4/

[46] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, *338*, 139518. https://doi.org/10.1016/j.chemosphere.2023.139518
[47] Saikiran, K., Lithesh, G., Srinivas, B., & Ashok, S. (2021). Prediction of Air Quality Index Using Supervised Machine Learning Algorithms. *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 1–4. https://doi.org/10.1109/ACCESS51619.2021.9563323
[48] Seo, D., Kim, Y., Eo, Y., Park, W., & Park, H. (2017). Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection. *Remote Sensing*, *9*(11), 1163. https://doi.org/10.3390/rs9111163
[49] Shataee, S., Kalbi, S., Fallah, A., & Pelz, D. (2012). Forest attribute imputation using machine-learning methods and ASTER data: Comparison of *k* -NN, SVR and random forest regression algorithms. *International Journal of Remote Sensing*, *33*(19), 6254–6280. https://doi.org/10.1080/01431161.2012.682661
[50] Sicard, P., Lesne, O., Alexandre, N., Mangin, A., & Collomp, R. (2011). Air quality trends and potential health effects – Development of an aggregate risk index. *Atmospheric Environment*, *45*(5), 1145–1153. https://doi.org/10.1016/j.atmosenv.2010.12.052
[51] Singh, A., N., M., & Lakshmiganthan, R. (2017). Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. *International Journal of Advanced Computer Science and Applications*, *8*(12). https://doi.org/10.14569/IJACSA.2017.081201
[52] Soni, H. B., & Patel, J. (2018). Assessment of Ambient Air Quality and Air Quality Index in Golden Corridor of Gujarat, India: A Case Study of Dahej Port. *International Journal of Environment*, *6*(4), 28–41. https://doi.org/10.3126/ije.v6i4.18908
[53] Teli, S., & Kanikar, P. (2015). A survey on decision tree-based approaches in data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 613–617.
[54] Upadhyay, S. (2020). Comparative Analysis of Machine Learning Regression Algorithms on Air Pollution Dataset. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 125–136. https://doi.org/10.32628/CSEIT206427
[55] Zhou, Y., De, S., Ewa, G., Perera, C., & Moessner, K. (2018). Data-Driven Air Quality Characterization for Urban Environments: A Case Study. *IEEE Access*, *6*, 77996–78006. https://doi.org/10.1109/ACCESS.2018.2884647
[56] Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., & Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. *Environmental Pollution*, *231*, 1232–1244. https://doi.org/10.1016/j.envpol.2017.08.069