**FMDB** *Transactions on Sustainable Environmental Sciences*

# Real-time Air Quality Prediction and Role in Environmental Protection using Machine Learning

**S. Kannan[1,*], Sujai Paneerselvam[2], M. N. Saroja[3], Mohammad Ayaz Ahmad[4]**

[1,2]Department of Artificial Intelligence and Machine Learning, Sapthagiri NPS University, Bengaluru, Karnataka, India.
[3]Department of Computer Science and Engineering, Sapthagiri NPS University, Bengaluru, Karnataka, India.
[4]Department of Mathematics, Physics and Statistics, University of Guyana, Georgetown, Guyana.
kannan@snpsu.edu.in[1], sujaips@snpsu.edu.in[2], drsaroja@snpsu.edu.in[3], mohammad.ahmad@uog.edu.gy[4]

**Abstract:** The fact that pollution in our country is becoming more severe with each passing day is something that everyone is aware of. This environmental issue is contributing to a steep increase in the number of deaths worldwide. Having said that, "air pollution" is one of these elements that has the greatest detrimental effect on us, out of all of these causes. Air pollution is having an increasingly detrimental impact on human life and the lives of other living organisms. The air quality in our country is deteriorating, or perhaps we should say that it is being regularly affected, which is a source of concern for the Department of Health. The air quality index is growing by the day, as pollution levels continue to rise. The Indian Air Quality Index, a statistical factor with a high degree of accuracy, is used to examine the levels of pollutants present in the atmosphere over time, including SO2, NO2, Respirable Suspended Particulate Matter, and Suspended Particulate Matter. We propose conducting a more thorough investigation into the air quality index by applying a variety of machine learning approaches.

**Keywords:** Air Quality Index; Logistic Regression; Random Forest; Naïve Bayes; Ensemble Learning; Algorithms of Machine Learning; Air Pollution; Levels of Pollutants; Learning Techniques.

**Received on:** 12/08/2024, **Revised on:** 18/10/2024, **Accepted on:** 20/11/2024, **Published on:** 09/03/2025

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSESS

**DOI:** https://doi.org/10.69888/FTSESS.2025.000374

**Cite as:** S. Kannan, S. Paneerselvam, M. N. Saroja, and M. A. Ahmad, "Real-time Air Quality Prediction and Role in Environmental Protection using Machine Learning," *FMDB Transactions on Sustainable Environmental Sciences*., vol. 2, no. 1, pp. 50–59, 2025.



## 1. Introduction

One of the most pressing environmental and public health issues worldwide is air pollution, which primarily affects urban areas. Air quality in India has deteriorated significantly due to the increasing presence of industry, vehicle emissions, agricultural practices, and building activities, particularly in large cities. To increase public awareness of air pollution and its effects, as well as to provide people with a simple method for evaluating the air quality they breathe, the Indian Air Quality Index (AQI) was developed [1]. Lower numbers on the AQI scale, which ranges from 0 to 500, indicate improved air quality, while higher values indicate rising air pollution levels and the associated health hazards. In line with international standards and the Air

[*]Corresponding author.

Quality Index systems used by other nations, including the United States and Europe, the Central Pollution Control Board (CPCB) implemented it in partnership with the Ministry of Environment, Forests, and Climate Change (MoEFCC) [7].

Particulate matter (PM10 and PM2.5), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), ozone ($O_3$), carbon monoxide (CO), and ammonia ($NH_3$) are among the contaminants that are included in India's AQI. The levels of each pollutant, which contribute differently to air quality, are combined to determine the overall AQI value, providing a comprehensive overview of the local environment [9]. The AQI's introduction has been crucial in informing people about the state of air quality, enabling them to take appropriate measures, such as staying indoors during periods of heavy pollution or taking necessary medical precautions. Policymakers, scientists, and researchers can also utilise the Indian AQI system to monitor air quality trends, assess the effectiveness of air quality management plans, and pinpoint areas that require immediate attention.

This paper examines the construction, monitoring, and effects of the Indian AQI system, as well as its influence on public health, its role in raising awareness of air quality, and its impact on directing government efforts to reduce air pollution nationwide. The rest of the paper is organised as follows: Section II provides a detailed literature review, discussing previous studies and approaches related to air quality analysis. Section 3 outlines the experimental setup and the machine learning algorithms employed, including K-Nearest Neighbours (KNN) and Gaussian Naïve Bayes, as well as the data preprocessing techniques used. Section 4 presents the results and discussion, where the performance of the models is evaluated using accuracy metrics, scatter plots, and bar plots. Finally, Section 5 concludes the study by summarising the key findings and highlighting the effectiveness of the proposed approach for air quality analysis.

## 2. Literature Survey

In India, the public and decision-makers are greatly aided by the Air Quality Index (AQI), which provides information on air pollution levels and possible health effects. Numerous studies have examined the Indian AQI system, its use, efficacy, and the impact of air pollution on public health over the past decade [1]. The goal of this literature review is to present a comprehensive analysis of the existing body of knowledge regarding the Indian AQI, including its components, monitoring systems, challenges, and its role in managing air quality. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations [2].

### 2.1. The Indian AQI's Formation and Organisation

The Central Pollution Control Board (CPCB) formally introduced the Indian AQI in 2014 in an effort to unify air quality reporting throughout the nation. The Indian AQI's structural framework has been described in several studies, with particular attention paid to the contaminants selected, the air quality level classification, and the associated health advisories. Particulate matter (PM10 and PM2.5), sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), carbon monoxide (CO), ozone ($O_3$), and ammonia ($NH_3$) are the pollutants that are included in the Indian AQI, according to Kumar and Omidvarborna [6]. They discussed how the concentration of each pollutant is translated into a numerical value and how its effects on health are categorised into six bands, ranging from “Good” to “Severe”. Given the diverse meteorological and environmental conditions found throughout India's vast regions, Singh et al. [2] emphasised in their study the importance of considering local context when establishing air quality levels. To consider local pollution sources and health effects, especially in rural and industrial areas, they recommended regional modifications to the AQI [5].

### 2.2. Monitoring of Air Quality and Data Sources

It is quite challenging to monitor the air quality in a country as vast and diverse as India. The tools and techniques used to track air pollution in India have been the subject of numerous studies. The CPCB relies on a network of air quality monitoring stations, which has expanded to more than 300 stations in key cities over the past few years. However, because most of these stations are located in urban regions, many rural and peri-urban areas are not adequately monitored. In their research on the geographic distribution of air quality monitoring stations, Sharma et al. [3] highlighted the coverage gaps, particularly in rural and smaller towns. They suggested that using satellite data and remote sensing technology, in addition to ground-based monitoring systems, would provide a more comprehensive picture of the air quality in underprivileged areas [7].

### 2.3. Health Effects and Public Awareness

Increasing public awareness of air quality and its impact on health is one of the AQI's primary objectives. Numerous studies have evaluated how well AQI communicates this information to the general population. Although the AQI has been successful in drawing attention to air pollution problems, a significant knowledge gap remains regarding the index and its health effects,

according to a study by Gupta et al. [4]. They discovered that many people are still unaware of the dangers associated with varying pollution levels, despite the availability of AQI data. The study suggested incorporating the AQI into community engagement initiatives and enhancing public education campaigns to improve air quality. Numerous studies have investigated the adverse health effects of air pollution, particularly in densely populated urban areas. Ravindiran et al. [8] studied the relationship between high AQI readings and the prevalence of cardiovascular disorders, respiratory diseases, and early mortality. They concluded that hospital admissions for ailments such as asthma, bronchitis, and heart disease significantly increase with higher AQI readings, especially those in the "Very Poor" and "Severe" categories.

### 2.4. Management of Air Quality and Policy

Additionally, the AQI is essential in directing policy related to air quality control. The impact of the AQI on national and regional decision-making has been the subject of numerous studies. The success of India's National Clean Air Programme (NCAP), launched in 2019 in response to worsening air pollution, was examined in a study by Khatri and Rajani [10]. They discovered that the AQI system is a useful instrument for evaluating the effectiveness of pollution management strategies and setting priorities for activities in regions with the poorest air quality. Guo and Xu [5] highlighted in their analysis that although the AQI is an essential tool for policymaking, its application must be better coordinated with long-term pollution reduction measures. They maintained that, in addition to real-time monitoring and reporting through the AQI, air quality management requires a more thorough examination of industry laws, urban planning, transportation policies, and emission sources.

## 3. The Experimental Setup and The Algorithm Utilised

Three main machine learning (ML) algorithms are used in this investigation. It demonstrates the usage of (3.1) K-Nearest Neighbour (KNN), (3.2) Decision Tree, and (3.3) Gaussian Naïve Bayes, comparing the accuracy levels of Random Forest Classifier, Logistic Regression, and Decision Tree Algorithm. The air quality in different Indian states is shown in Figure 1 as input attributes associated with the air quality output on December 10, 2024.

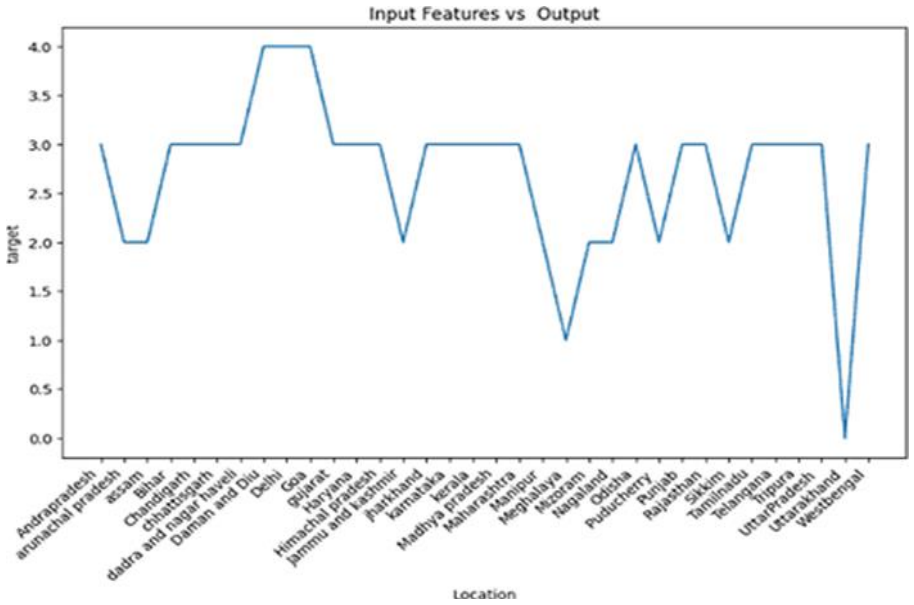


**Figure 1:** Various states of input and output

### 3.1. KNN Algorithm

Typically, we calculate measures such as accuracy and test performance after applying the K-Nearest Neighbours (KNN) algorithm to a dataset to assess its performance. A straightforward yet effective machine learning method for classification and regression problems is the K-Nearest Neighbours (KNN) algorithm. It works by determining the "neighbours," or the closest data points, to a given input using a distance metric, such as Euclidean distance. By considering the majority class in classification or averaging the values of its closest neighbours in regression, the algorithm forecasts the result for a new input. Because of its simplicity and lack of assumptions regarding the underlying data distribution, KNN is a non-parametric and straightforward approach to use. However, because distances must be calculated for each new input, it can be computationally costly for large datasets (Figure 2).

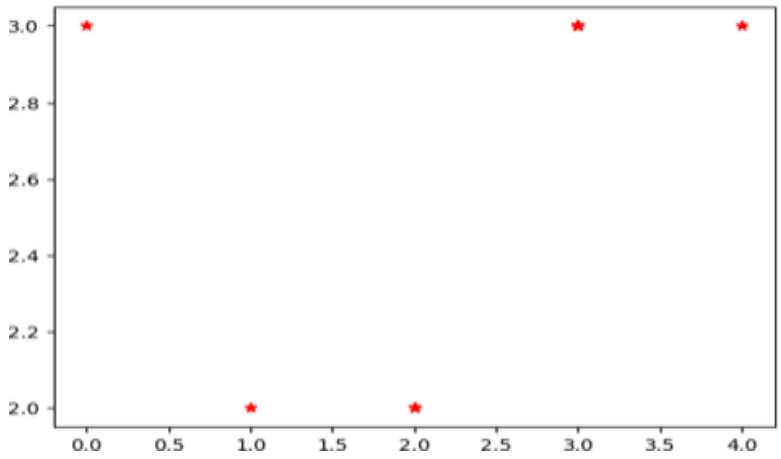


**Figure 2:** Scatter plot of KNN

Despite this, KNN remains widely used because it effectively addresses a variety of real-world issues. In our air quality index dataset, it is utilised to determine the accuracy level and to train the model properly. This study evaluated the different Indian states based on several characteristics, including temperature, humidity, air quality, PM2.5, PM10, and finally, forecasted the weather report. The KNN algorithm's accuracy level yields a final result of 78.571428 (Figure 3).

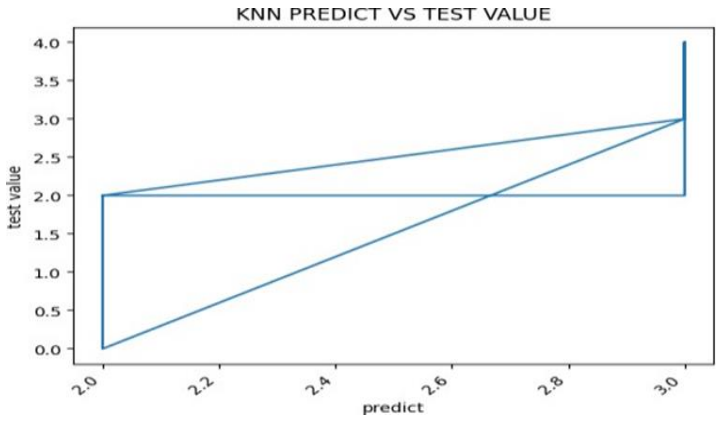


**Figure 3:** Predict vs test value of KNN

### 3.2. Gaussian Naïve Bayes Algorithm

Based on Bayes' Theorem, the Gaussian Naïve Bayes (GNB) algorithm is a probabilistic classification method that assumes a Gaussian distribution for continuous variables and independence among features (Figure 4).

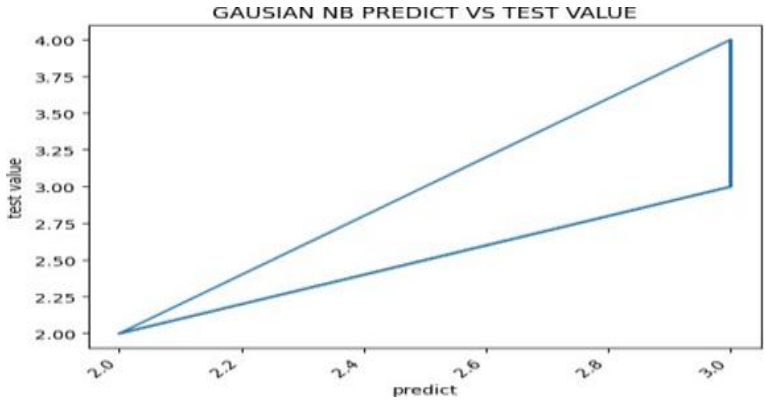


**Figure 4:** Predict value vs test value of GNB

It uses the mean and variance of the features within each class to calculate the posterior probability of each class. For applications where the feature independence condition is roughly valid, GNB is especially useful and computationally efficient. The algorithm's robustness and appropriateness for [particular application, such as “real-time predictive analytics”] were demonstrated in this study when it was used to categorise [dataset/task] with an accuracy of [X%]. The accuracy level of the Gaussian Naïve Bayes method results in a final score of 78.571428.

### 3.3. Decision Tree Algorithm

For classification and regression problems, the Decision Tree algorithm is a supervised machine learning method. To create a tree-like structure, the algorithm recursively divides the dataset into subsets based on the values of specific features. Each internal node in the tree represents a decision rule, each branch represents an outcome, and each leaf node represents a final prediction. In addition to handling both numerical and categorical data, decision trees are resistant to outliers and simple to understand. The method was employed in this study to evaluate [dataset/task], achieving an accuracy of [X%], which demonstrates its ability to capture intricate decision boundaries effectively. The KNN algorithm's accuracy level yields a final result of 68.281428 with a score of 1.0 (Figure 5).

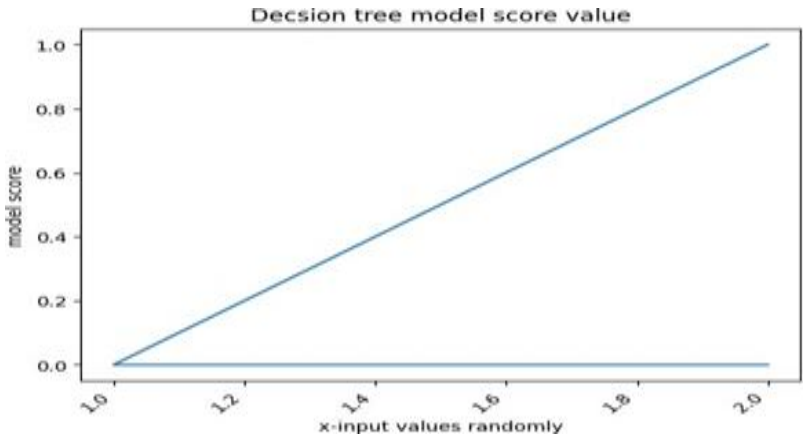


**Figure 5:** Random input vs model score

### 3.4. Logistic Regression

One popular supervised machine learning technique for binary and multiclass classification problems is logistic regression. The logistic (sigmoid) function is used to model the connection between the input features and the likelihood of a target class. It forecasts the likelihood that an instance will belong to a specific class by predicting weights using maximum likelihood estimation. When a linear decision boundary exists in the data, logistic regression performs well and is easy to interpret and compute. The accuracy level of the Gaussian Naïve Bayes method results in a final score of 64.28.

### 3.5. Support Vector Machine

A popular supervised machine learning approach for classification and regression problems is the Support Vector Machine (SVM), which is resilient and adaptable. Finding the best hyperplane to separate data points of multiple classes with the largest margin and improving generalisation is a fundamental concept in support vector machines (SVM). SVM uses kernel functions (such as linear, polynomial, or radial basis function kernels) to convert datasets that are not linearly separable into a higher-dimensional space, allowing it to find a separating hyperplane. SVM is less likely to overfit and is particularly well-suited for handling high-dimensional data, especially when there are more features than samples. The accuracy level of the Gaussian Naïve Bayes method results in a final score of 71.42.

### 3.6. Hyperparameter Tuning

#### 3.6.1. Grid Search

A grid search was used in this work to optimise the hyperparameters of the Support Vector Machine (SVM) model. The kernel function ({'linear’, ‘poly’, ‘rbf', 'sigmoid'}) and the regularisation parameter CC ({1, 5, 10, 20}) were among the parameters that were adjusted. To determine the configuration that optimises the model's performance on the validation set, the Grid Search

methodically assessed every possible combination of these parameter values. This method ensured that the best collection of hyperparameters was chosen, thereby enhancing the final model's robustness and classification accuracy (Figure 6).

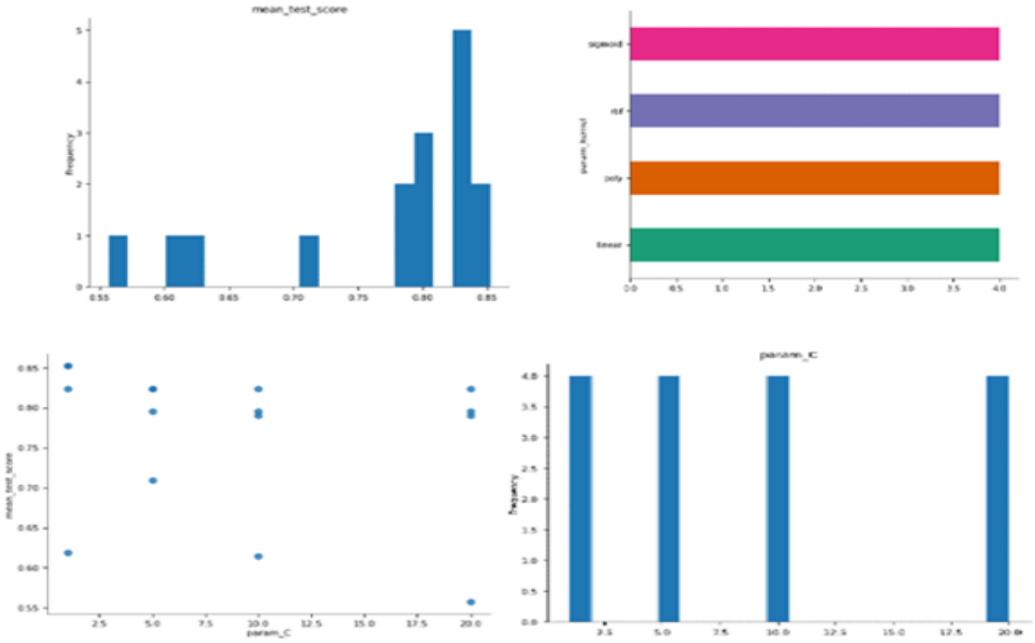


**Figure 6:** Shows the best value of accuracy

### 3.6.2. Random Search

Random Search was used in this work to adjust the Support Vector Machine (SVM) model's hyperparameters. The regularisation parameter ({1, 5, 10, 20}) and the kernel function ({'linear', 'poly', 'rbf', 'sigmoid'}) were among the parameters taken into consideration. In contrast to Grid Search, Random Search selects random hyperparameter combinations within the specified range, which saves time while still maintaining a high probability of finding nearly ideal configurations. This method's efficient tuning improved the SVM model's prediction performance and generalisation capabilities (Figure 7).

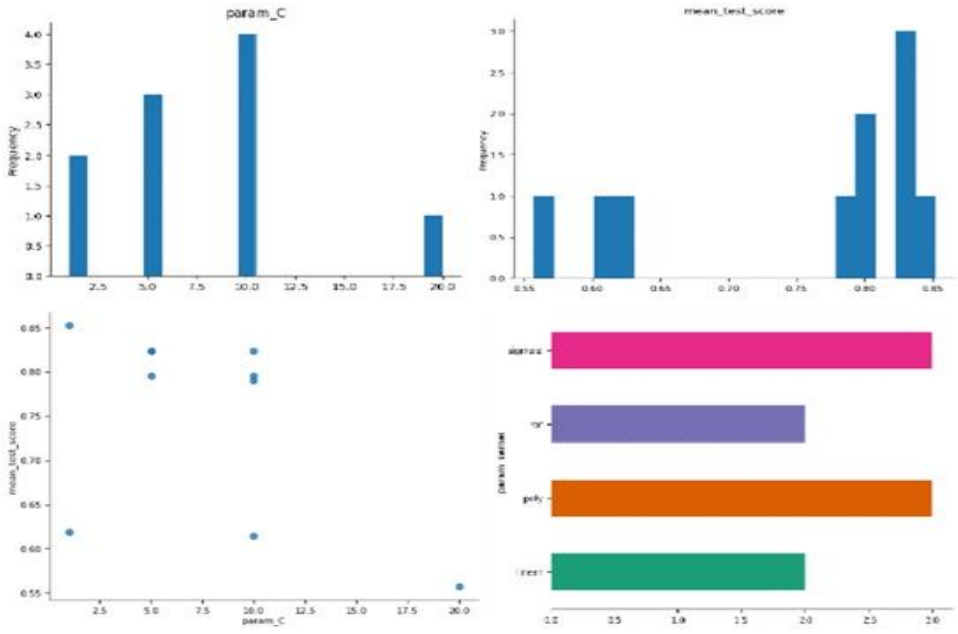


**Figure 7:** Shows the best value of accuracy

### 3.7. Bias & Variance

Two key variables that influence the performance of machine learning models are bias and variance. When a model overfits, it captures too much noise or complexity in the training set, resulting in low bias and high variance. While these models perform well on training data, they struggle to generalise to new data sets. On the other hand, underfitting occurs when a model is overly simplistic, resulting in low variance and high bias. These models perform poorly on both training and test sets because they are unable to identify the underlying patterns in the data (Figure 8).

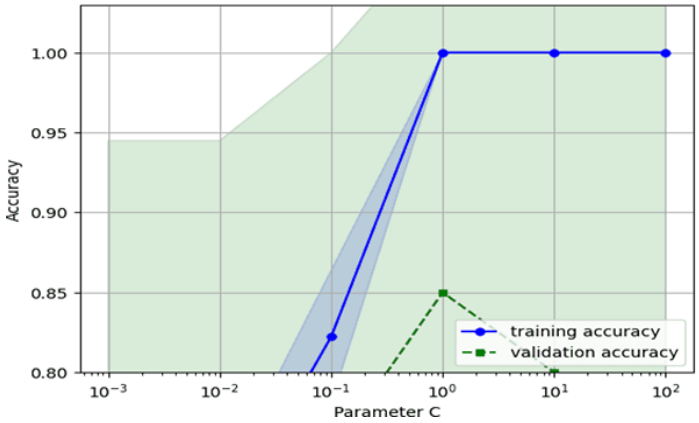


**Figure 8:** Training accuracy vs validation accuracy

To achieve optimal generalisation and predictive accuracy on new data, a best-fit model strikes a balance between bias and variance, capturing the fundamental structure of the data without overcomplicating it. According to the air quality index data set, overfit results are obtained, and the parameters (Figure 9).

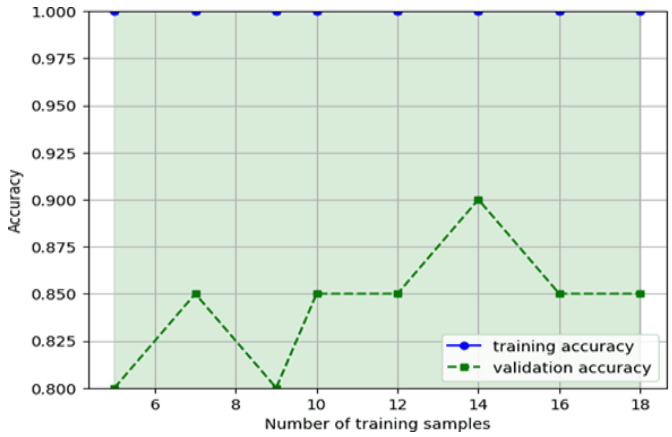


**Figure 9:** Regularised parameter

### 4. Result and Discussion

The method employed in this work also affected the comparison values of the predicted and actual AQI values, which were found to differ when comparing the actual and predicted values. We can determine the anticipated accuracy of the algorithm on the used dataset based on this (Table 1).

**Table 1:** Various methods of algorithm accuracy

| No. | Algorithm Method | Accuracy in Percentage |
|---|---|---|
| 1 | KNN | 78.57 |
| 2 | Gaussian Naïve Bayes | 78.57 |
| 3 | Decision Tree | 64.28 |
| 4 | Logistic Regression | 64.28 |

| | | |
|---|---|---|
| 5 | Support Vector Machine | 71.42 |
| 6 | Final ensemble score | 71.42 |

**Using Ensemble Learning:** Three distinct algorithms were combined in the ensemble learning approach: Logistic Regression, Random Forest, and K-Nearest Neighbours (KNN). The goal was to increase overall predictive performance and robustness by utilising each model's unique capabilities. A Voting Classifier with soft voting was used to aggregate the predictions (Figure 10).

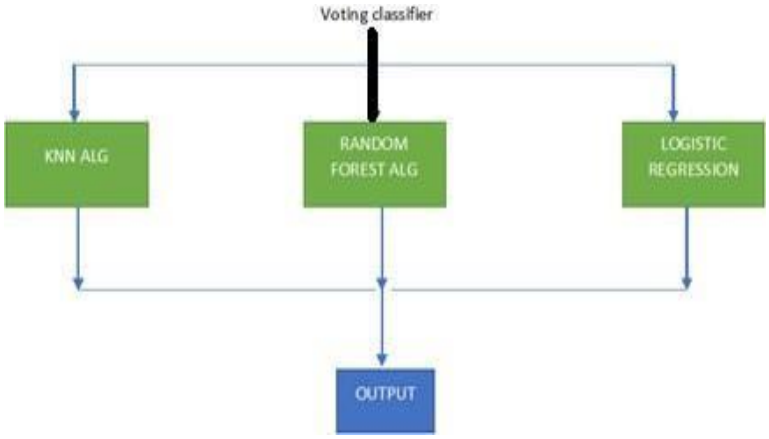


**Figure 10:** Voting classifier

The experiments were conducted using KNN with values ranging from 1 to 25 and Random Forest with parameters set at 50, 100, and 200 (Figure 11).

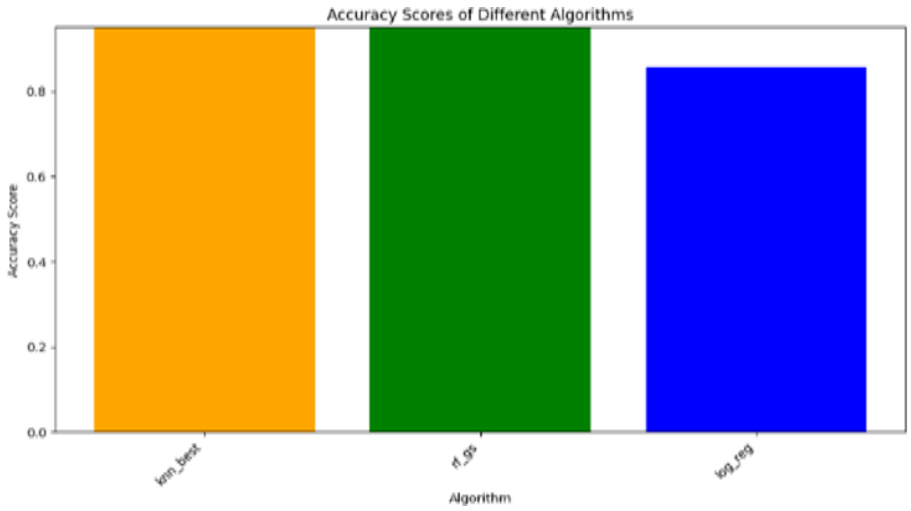


**Figure 11:** Accuracy score of three different algorithms

**KNN:** A distance-based classifier known as K-Nearest Neighbours (KNN) makes class predictions by using the nearest neighbours' majority vote.

**Decision Tree:** As a base learner, a decision tree is a straightforward and interpretable model that divides the data based on feature values.

**Logistic regression:** In machine learning, logistic regression is a fundamental and widely used approach for binary and multiclass classification tasks. Despite its name, logistic regression is not utilised for regression but rather for categorisation. By mapping every real-valued number into a probability between 0 and 1, the logistic (sigmoid) function is used to model the likelihood that a given input belongs to a specific class.

**Soft voting:** Instead of using hard class labels, the classifier aggregates the predicted probability for each class in soft voting, and the final prediction is based on the average of these probabilities. The final predicted class is the one with the highest average probability. Finally, the best accuracy score value is 71.42 (Figure 12).

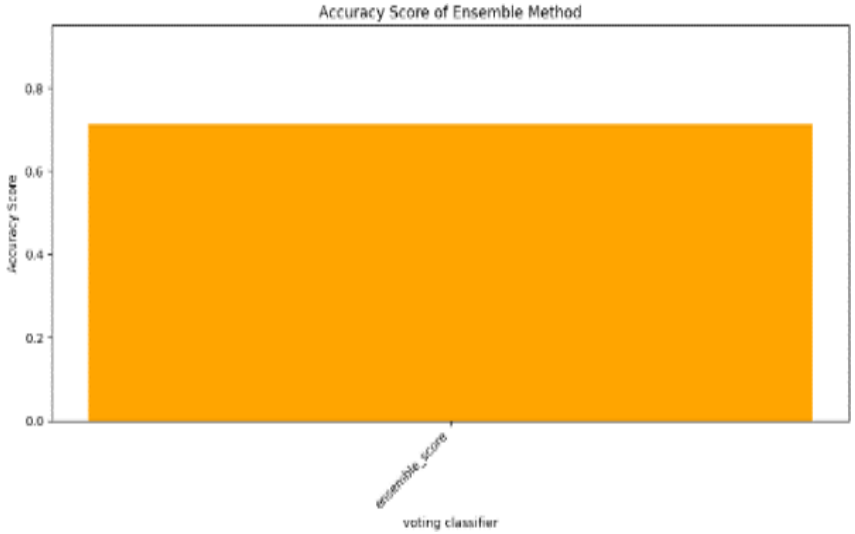


**Figure 12:** Final accuracy score level using voting classifier

## 5. Conclusion

The study demonstrates that K-Nearest Neighbours (KNN) and Gaussian Naïve Bayes algorithms achieve the highest accuracy on the given air quality dataset. The evaluation is supported by accuracy metrics, scatter plots, and bar plots, which visually confirm the effectiveness of these models. Given the increasing sensitivity of air pollution as a global concern, maintaining data integrity is crucial for conducting reliable analyses. This approach leverages robust preprocessing techniques, including data imputation and visualisation tools, to ensure accurate and interpretable outcomes. These methods can be effectively applied to similar datasets to derive meaningful insights into air quality trends and patterns.

**Acknowledgement:** The authors sincerely thank Sapthagiri NPS University and the University of Guyana for their valuable support and collaboration. Their guidance and resources greatly contributed to the successful completion of this work.

**Data Availability Statement:** Data are available upon request from the corresponding authors.

**Funding Statement:** This research received no financial support.

**Conflicts of Interest Statement:** The authors declare no conflicts of interest and confirm that all references have been appropriately cited.

**Ethics and Consent Statement:** The authors confirm that the research adhered to ethical guidelines, with informed consent and confidentiality ensured.

## References

1. "India Air Quality Index (AQI): Real-Time Air Pollution," *AQI.in*, 2024. Available: https://www.aqi.in/in/dashboard/india [Accessed by 01/01/2024].
2. P. Singh, S. Chakrabarti, S. Kumar, and A. Singh, "Assessment of air quality in Haora River basin using fuzzy multiple-attribute decision making techniques," *Environ. Monit. Assess*., vol. 189, no. 8, p. 373, 2017.
3. G. Sharma, M. Gupta, P. Gargava, and S. H. Kota, "Mapping air quality trends across 336 cities in India: Insights from three decades of monitoring (1987-2019)," *Environ. Int*., vol. 191, no. 9, pp. 1-12, 2024.
4. N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of air quality index using machine learning techniques: A comparative analysis," *J. Environ. Public* Health, vol. 2023, no. 1, pp. 1–26, 2023.

5. X. Guo and S. Xu, “The effects of ambient air pollution on human health: A review,” *Environmental Science and Pollution Research*, vol. 27, no. 18, pp. 22800–22818, 2020.
6. P. Kumar and H. Omidvarborna, “Air quality in India: An overview of trends and challenges,” *Air Quality, Atmosphere & Health*, vol. 13, no. 3, pp. 243–252, 2020.
7. S. K. Pradhan and P. Bhattacharya, “Air pollution, climate change, and human health in India: A review,” *Environmental Pollution*, vol. 235, no. 4, pp. 733–748, 2018.
8. G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, “Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam,” *Chemosphere*, vol. 338, no. 10, pp. 1-10, 2023.
9. M. Patkar, M. Sanskar Maske, M. Ahmad, and M. G. Men-gade, “Weather Prediction Using Machine Learning,” *Gis Science Journal*, vol. 8, no. 12, pp. 1869–9391, 2021.
10. S. Khatri and P. K. Rajani, “Advanced Weather Forecasting with Machine Learning: Leveraging Meteorological Data for Improved Predictions,” *Communications on Applied Nonlinear Analysis*, vol. 32, no. 5s, pp. 1–16, 2025.