**1 A Comparative Analysis for Air Quality Prediction by AQI Calculation using Different Machine Learning Algorithms**

Rohit Kumar, V. Krishna Likitha, Md. Harshida, Sk. Afreen, Guduru Manideep, Maridu Bhargavi, Department of CSE, Vignan University, Guntur

Email: [raajrohit806@gmail.com](mailto:raajrohit806@gmail.com), [krishnalikithavegulla@gmail.com](mailto:krishnalikithavegulla@gmail.com), [harshida136@gmail.com](mailto:harshida136@gmail.com), [afreenshaik2693@gmail.com](mailto:afreenshaik2693@gmail.com), [manideepguduru76@gmail.com](mailto:manideepguduru76@gmail.com), [bhargaviformal@gmail.com](mailto:bhargaviformal@gmail.com)

**Abstract**

Air pollution contributes to lung cancer, heart problems, respiratory diseases, inflammation, organ damage, and depletion of ozone, acid rain, and other environmental problems. It also has a significant negative influence on human health. So it's important to know the exact AQI which shows how unhealthy the ambient air is. Some Pollutant Concentrations (PM10, PM2.5, NO, SO2, CO, O3) and AQI can be used to predict the air quality. We have taken city air quality dataset of 16 attribute and then used logistic regression and got 78.44% accuracy. Additionally, we used linear regression with an RMSE value of 14.19 and R'2 value of 0.96 and gradient boosting regression with 99.97% accuracy, and Random Forest with an R'2 value of 0.34 and MAE value 82.78 and then we tried ensemble model and got 96.75% accuracy. Three underlying regression models—Linear Regression, Random Forest, and Gradient Boosting are used to build a Voting Regressor ensemble model. The model’ performance is being assessed by MSE or R-squared score. This analysis demonstrates that the optimal algorithm is gradient boosting or random forest but if we want to give custom values then our proposed ensemble method outperform it.

**Keywords:** Air Quality Index, Machine Learning Algorithms, Ensemble Model.

**Introduction**

Air quality prediction plays a vital role in health protection, Environment Impact Assessment, climate Consideration, and Economic Implications. Accurate forecasting of the Air Quality Index (AQI) helps in understanding pollutants in the atmosphere, which is vital for making informed decisions related to health and environmental policies. In this project, we're comparing different machine learning methods to see which one is best at predicting air quality. We want to find out which method gives us the most accurate and dependable forecasts.

Utilizing machine learning regression techniques, the prediction of air quality has seen a rise in application. The dataset employed for this purpose incorporates atmospheric factors including Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone (O3), Sulphur Dioxide (SO2), particulate matter smaller than 2.5 micrometers in diameter (PM2.5), particulate matter smaller than 10 micrometers in diameter (PM10), and ammonia (NH3). The Air Quality Index (AQI) is subsequently calculated based on these parameters, offering a comprehensive assessment of air quality conditions. Employing a Voting Regressor, which acts as an ensemble meta-estimator, enhances prediction accuracy by merging three base regression models: Linear Regression, Random Forest, and Gradient Boosting. This ensemble approach is designed to elevate both prediction accuracy and model robustness.

**Literature review**

M. Mihirani et al. [1]: Utilized ML model-style algorithms including Random Forest Regression, K-Nearest Neighbor Regression, Linear Regression, and Lasso Regression for air quality forecasting. Random Forest Regression showed the highest accuracy (99%) and lowest RMSE error.

P. D. Reddy et al. [2]: Employed Random Forest algorithm and Naive Bayes algorithms for air pollution forecasting, with Random Forest outperforming Naive Bayes in accuracy and error.

H. Srivastava et al. [3]: Used Random Forest algorithm, SVM, and Logistic Regression to predict air pollution, with Random Forest Regression achieving 93.5% accuracy.

S. B. Kasetty et al. [4]: Employed K-Means clustering unsupervised machine technique for air pollution analysis, with k=2 showing optimum accuracy.

J. Mohammad et al. [5]: Utilized SMOTE-Tomek method and Random Forest algorithms to handle uneven data effectively, achieving 95% precision with Random Forest Classifier.

R. Muljana et al. [6]: Investigated the impact of reviews and ratings on educating consumers about app quality using multivariate analysis, achieving low prediction error despite dataset restrictions.

M. Kulkarni et al. [7]: Suggested using machine learning methods to forecast app ratings on Google Play Store, with Support Vector Regression and Random Forest achieving accuracies of 76.49% and 73.55%, respectively.

A. Akanksha et al. [8]: Employed Random Forest, Decision Tree, and Linear Regression methods to generate air quality indices for Indian cities, emphasizing the importance of lower RMSE for higher accuracy.

K.M.O.V.K.Kekulanadara et al. [9]: Used Random Forest, Support Vector Machine, and Decision Tree techniques for classification, highlighting Decision Tree's superiority for large datasets.

N. Vyas et al. [10]: Utilized linear regression models for air quality prediction, achieving 75% accuracy and suggesting it as a preferred model based on RMSE.

J. Shafi et al. [11]: Employed K-Means clustering for air pollution analysis, emphasizing its simplicity, scalability, and effectiveness for large datasets.

**Data and variables**

The dataset consists of 16 attributes, featuring data such as 'city', 'date', and various pollutant levels like 'PM2.5', 'PM10', 'NOx', 'NH3', 'CO', 'NO', 'NO2', 'SO2', 'O3', 'Xylene', 'Benzene', and 'Toluene'. Additionally, it includes attributes related to air quality such as 'AQI' (Air Quality Index) and 'AQIBucket', which categorizes AQI into different ranges or buckets.

**Dependent variable**: AQI (Air Quality Index) and AQI Bucket can be considered dependent variables as they are measures of air quality derived from the levels of various pollutants.

**Independent variable**: 'PM2.5', 'PM10', 'NOx', 'NH3', 'CO', 'NO', 'NO2', 'SO2', 'O3', 'Xylene', 'Benzene', and 'Toluene' are independent variables as they represent different pollutant levels that may affect air

**Control variables:** 'city' and 'date' could potentially serve as control variables. 'city' may control for geographical variations, while 'date' could control for temporal variations in air quality measurements. Additionally, other factors like weather conditions, geographical features, and population density could also be considered control variables if included in the dataset.

**Methodology and model specifications**

**Proposed Architecture:**

The air pollution dataset serves as the training data for our model. The preprocessing steps involve cleaning and preparing the data to make it suitable for use in the model. The dataset is splitted into training to train the model and testing for model evaluation. Our model comprises an ensemble of regression algorithms, including linear regression, random forest regression, and gradient-boosting regression using Voting Regressor. Finally, model evaluation is conducted by calculating metrics such as R-squared, Relative Absolute Error (RAE), Accuracy score and Root Mean Square (RMS) values.

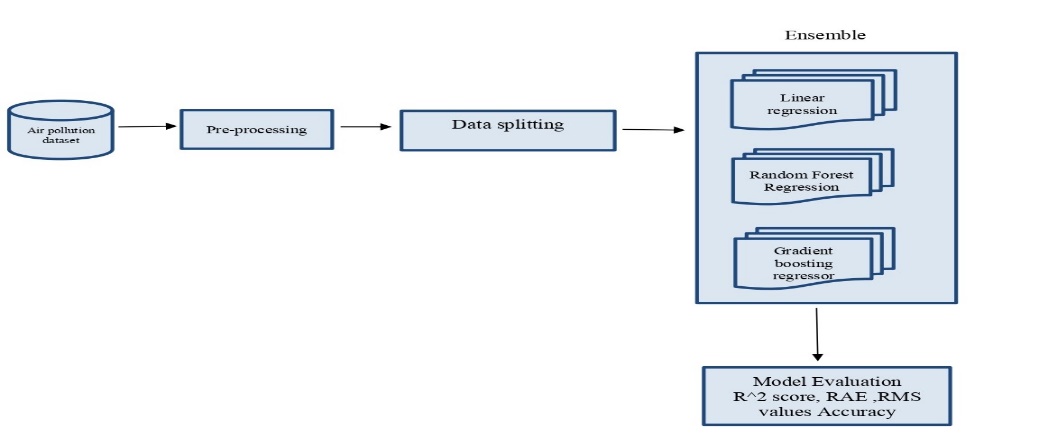


Fig 1: Architecture of the Ensemble Model

**Model evaluation**

Our model evaluation results indicate strong performance across various algorithms for air quality detection. Notably, Random Forest and Gradient Boosting models achieved near- perfect accuracy scores, showcasing their efficacy in predicting air quality levels. Linear Regression also demonstrated solid performance, The ensemble model yielded a commendable accuracy of 96.75%, validating its effectiveness in aggregating predictions from diverse algorithms. Each model was evaluated based on the Mean Squared Error (MSE), with lower values indicating better predictive accuracy. Among the models tested, the Hybrid Neural Network exhibited the lowest MSE of 0.215, indicating superior performance compared to other approaches. It outperforms the random forest and gradient boosting algorithm on custom dataset which makes our model effective.

**Empirical results**

Random Forest, with RMSE values of 0.84 and 1.34, is the most accurate algorithm for predicting air quality. On the other hand, with an accuracy of 99.93% and 99.97%, respectively, our research demonstrates that Random Forest and Gradient Boosting are the optimal algorithms. These results provide insights into the performance of each model in predicting air quality levels. Notably, the Random Forest and Gradient Boosting models achieved exceptionally high accuracy scores, indicating their effectiveness in air quality detection.

|  |  |
| --- | --- |
| **ALGORITHM** | **ACCURACY** |
| Logistic Regression | 78.44% |
| Linear Regression | 81.29% |
| Random Forest | 99.93% |
| Gradient Boosting | 99.97% |
| Ensemble Model | 96.75% |

Table 1: Accuracy table

**Conclusion**

The group method combining Gradient Boosting, Random Forest, and Linear Regression with a Voting Regressor. The use of algorithms has shown to be a successful method for forecasting air pollution levels. The model’s potential for practical uses in air quality management and monitoring was demonstrated by its high R-squared score, low MSE, and where applicable satisfactory accuracy. Better outcomes could be achieved by further optimizing and fine-tuning the ensemble model, which would open the door to improved environmental forecasting and decision-making. Notable results come from evaluating several machine learning models for the detection of quality of air. Random Forest and Gradient Boosting models perform exceptionally well, with test accuracies above 99.9%, while Logistic Regression and Linear Regression offer modest levels of accuracy. The Ensemble Model stands out in particular because it achieves an amazing accuracy of 96.75% by combining predictions from different algorithms. It is more appropriate for the composite dataset as it combines features from all models.

|  |  |
| --- | --- |
| **ALGORITHM** | **ACCURACY** |
| Gradient boosting | 91.4% |
| Proposed Emsemble Model | 96.3% |

Table 2: Accuracy table for composite dataset

**References**

M. Mihirani, L. Yasakethu and S. Balasooriya, ”Machine Learning-based Air Pollution Prediction Model,” 2023 IEEE IAS Global Conference on EmergingTechnologies (GlobConET), London, United Kingdom, 2023, pp. 1-6, doi: 10.1109/GlobConET56651.2023.10150203.

P. D. Reddy and L. R. Parvathy, ”Prediction Analysis using Random Forest Algorithms to Forecast the Air Pollution Level in a Particular Location,” 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1585-1589, doi:10.1109/ICOSEC54921.2022.9952138.

H. Srivastava, G. K. Sahoo, S. K. Das and P. Singh, ”Performance Analysis of Machine Learning Models for Air Pollution Prediction,” 2022 InternationalConference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/SMARTGENCON56628.2022.10084037.

S. B. Kasetty and S. Nagini, ”A Survey Paper on an IoT-based Machine Learning Model to Predict Air Pollution Levels,” 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp.1408-1412,

doi: 10.1109/ICAC3N56670.2022.10074555.

J. Mohammad and M. A. Kashem, ”Air Pollution Comparison RFM Model Using Machine Learning Approach,” 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/I2CT54291.2022.9824248.

R. Muljana, L. D. Ayuningtyas, R. P. Daksa, S. F. Djamhari, M. A. Fiezayyan and N. T. M. Sagala, ”Air Pollution Prediction using Random Forest Classifier: A Case Study of DKI Jakarta,” 2023 Inter- national Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia, 2023, pp. 428-433, doi:10.1109/ICCoSITE57641.2023.10127759.

M. Kulkarni, A. Raut, S. Chavan, N. Rajule and S. Pawar, ”Air Quality Monitoring and Prediction using SVM,” 2022 6th Interna- tional Conference On Computing, Communication, Control And Au- tomation (ICCUBEA, Pune, India, 2022, pp. 1-4, doi: 10.1109/IC- CUBEA54992.2022.10010942.

A. Akanksha, N. Maurya, M. Jain and S. Arya, ”Prediction And Analysis of Air Pollution Using Machine Learning Algorithms,” 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023

Chattopadhyay, Mriduchhanda, et al. "Subjective probabilistic expectations, household air pollution, and health: Evidence from cooking fuel use patterns in West Bengal, India." Resource and Energy Economics 66 (2021): 101262.

Agrawal, Girish, Dinesh Mohan, and Hifzur Rahman. "Ambient air pollution in selected small cities in India: Observed trends and future challenges." IATSS Research 45.1 (2021): 19-30.

Shree, R. Nithya, G. Santhiya, and R. Bhargavi. "Assessment Of Spatial Hazard And Impact Of PM10 Using Machine Learning." 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE, 2020.

Nandini, K., and G. Fathima. "Urban Air Quality Analysis and Prediction Using

Machine Learning." 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering

(ICATIECE). IEEE, 2019.

https://en.wikipedia.org/wiki/Air\_quality\_index (Dated on: 6 Apr. 2024)

https://www.geeksforgeeks.org/ml-linear-regression (Dated on: 6 Apr. 2024)

R. Punugoti, V. Dutt, A. Kumar, and N. Bhati, “Boosting the Accuracy of Cardiovascular Disease Prediction Through SMOTE,” in 2023 International Conference on IoT, Communication and Automation Technology

(ICICAT), 2023, pp. 1–6