International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 03 | Mar 2025 www.irjet.net p-ISSN: 2395-0072

Indian Air Quality Forecasting using Ensemble Methods for Early Warning Systems

Manpreet Hire[1], Pritesh Yadav[2], Smriti Pandey[3], Priyanshu Prajapati[4]

[1]*Asst. Professor, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[2]*Student, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[3]*Student, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[4]*Machine Learning Engineer Intern, BNK Infotech Pvt. Ltd., Delhi, India*

Abstract - *Air pollution remains a significant challenge in India, affecting public health, ecosystems, and economic growth. Accurate air quality forecasting is crucial for mitigating the adverse effects of pollution and enabling timely interventions. This study presents an ensemble-based machine learning approach to predict air quality across major Indian cities using historical air quality data. Various meteorological and pollutant concentration features are utilized to enhance prediction accuracy. The proposed methodology integrates multiple machine learning models, including Random Forest, Gradient Boosting, and XGBoost, to leverage their individual strengths and reduce forecasting errors. The ensemble framework is optimized to improve generalization and robustness in highly dynamic environmental conditions. The system is further designed to support early warning mechanisms, providing authorities and the public with timely alerts. Experimental results indicate that ensemble learning significantly enhances forecasting precision, making it a reliable tool for air quality management in India. This study underscores the potential of machine learning-driven forecasting in environmental monitoring and highlights the need for data-driven policy decisions to combat air pollution effectively.*

Key Words: Air Quality Forecasting, Ensemble Learning, Machine Learning, Early Warning Systems, Air Pollution Prediction, Environmental Monitoring, Random Forest, Gradient Boosting, XGBoost, Meteorological Data, Pollutant Concentration, India, Data-Driven Policy, Predictive Modeling, Public Health, Air Quality Index (AQI).

1. INTRODUCTION

Air pollution is a critical environmental and public health concern in India, with rising pollutant levels posing severe risks to human well-being and ecological balance. Rapid urbanization, industrialization, and vehicular emissions have exacerbated air quality deterioration, necessitating robust forecasting systems to enable timely interventions. Traditional air quality prediction models often struggle with the complex and dynamic nature of pollution patterns, making machine learning (ML) an effective alternative for enhancing predictive accuracy.

This study employs an ensemble learning approach to forecast air quality using data collected from control stations across multiple Indian cities spanning various states from 2010 to 2023. Ensemble learning, which integrates multiple base learners to improve prediction performance, is particularly well-suited for handling the high variability and non-linearity of air pollution data. By leveraging advanced ML techniques such as Random Forest, Gradient Boosting, and XGBoost, the proposed model aims to provide precise and reliable air quality predictions. These ensemble models collectively enhance generalization, minimize errors, and offer a robust framework for forecasting air pollution levels under diverse environmental conditions.

1.1 BACKGROUND

India has been grappling with severe air pollution for over a decade, with major metropolitan regions frequently experiencing hazardous air quality levels. Various pollutants, including PM2.5, PM10, NO_2, SO_2, and CO, contribute to deteriorating atmospheric conditions. Existing forecasting models often fail to capture the intricate relationships between meteorological parameters and pollutant concentrations. Machine learning, particularly ensemble methods, has emerged as a powerful tool for overcoming these limitations by combining multiple predictive models to improve accuracy and robustness. This study harnesses past air quality data from multiple Indian states to develop an ensemble-based early warning system, aiding in proactive pollution management.

1.2 TOOLS & TECHNOLOGIES

This study utilizes Python as the primary programming language due to its versatility in data science and machine learning applications. For implementing ensemble learning techniques, Scikit-Learn is employed, offering robust algorithms such as Random Forest and Gradient Boosting, which improve prediction accuracy by aggregating multiple weak learners. Additionally, XGBoost (Extreme Gradient Boosting) is integrated for its efficiency in handling large datasets and superior performance in predictive modeling. The dataset, collected from air

quality control stations across various Indian states from 2010 to 2023, undergoes preprocessing using Pandas and NumPy, ensuring effective data manipulation. Visualization and exploratory data analysis are conducted using Matplotlib and Seaborn, allowing for a deeper understanding of pollution trends and feature correlations. The machine learning pipeline, from data preprocessing to model training and evaluation, is implemented in Jupyter Notebook, ensuring an interactive and reproducible workflow. Performance evaluation metrics such as RMSE, MAE, and R^2 are used to assess model accuracy. The integration of these tools and technologies enables the development of a scalable and high-performance forecasting system, supporting early warning mechanisms for proactive air quality management in India.

2. LITERATURE REVIEW

Sr. No.	Title	Objectives	Findings
1.	**Air Quality Prediction using Ensemble Machine Learning Algorithm**	Compare regression models to estimate AQI based on six air pollutants.	The stacking method outperformed linear regression, k-NN regression, and decision tree regression in AQI estimation.
2.	**Forecasting Air Quality in India through an Ensemble Clustering Technique**	Analyze air quality data using ensemble clustering to predict trends.	The modified consensus function improved clustering performance, aiding in understanding pollutant effects across Indian cities.
3.	**Study and Development of Hybrid and Ensemble Forecasting Models for Air Quality Index Forecasting**	Develop a hybrid model combining ARFIMA and SVM for AQI forecasting.	The Additive-ARFIMA-SVM model with functionally expanded inputs improved AQI forecasting performance by 16.34% compared to ARIMA.
4.	**An Ensemble Deep Learning Model for Forecasting Hourly $PM_{2.5}$ Concentrations**	Propose an ensemble deep learning model to predict hourly $PM_{2.5}$ concentrations.	The model achieved a root mean squared error of 12.96 μg/m^3 for 24-hour ahead $PM_{2.5}$ forecasting, outperforming various deep learning models.
5.	**Can Artificial Rain, Drones, or Satellites Clean Toxic Air?**	Explore technological interventions for air pollution control in India.	Technologies like drone monitoring and satellite data can identify pollution sources, but effective enforcement and data-driven policies are essential for long-term improvement.
6.	**Ambient Air Quality Assessment Using Ensemble Techniques**	Assess air quality by applying ensemble techniques to classify pollution levels.	The random forest classifier demonstrated high accuracy in predicting air quality categories, suggesting its effectiveness for air quality assessment.
7.	**AirNet: Predictive Machine Learning Model for Air Quality Forecasting**	Develop a predictive machine learning model for air quality forecasting.	The proposed model demonstrated high accuracy in predicting air quality, highlighting the potential of machine learning approaches.

3. ALGORITHMS/TECHNIQUES

3.1. RANDOM FOREST REGRESSOR

A random forest regression model combines multiple decision trees into a single ensemble. Each tree is built from a different subset of the data and makes an independent prediction. The final output is determined by averaging or taking a weighted average of all the trees' predictions. Each tree makes its own prediction, and the model aggregates these predictions to generate a final result. In regression tasks, random forest predicts continuous target variables, reducing variance and improving accuracy by combining the outputs of several trees.[1]

Random Forest Regressor is particularly suitable for air quality forecasting due to its robustness against overfitting and ability to handle high-dimensional meteorological and pollutant data. By constructing an ensemble of decision trees trained on bootstrapped samples with random feature selection, it captures complex patterns across different atmospheric layers and emission sources.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

Where:

- $\hat{y}$ = Predicted AQ,
- N = Total number of trees,
- $h_i(x)$ = Output of i-th tree for input vector x

3.2. GRADIENT BOOSTING REGRESSOR

Gradient Boosting is an ensemble machine learning technique that builds a series of decision trees, each aimed at correcting the errors of the previous ones. For regression tasks, Gradient Boosting adds trees one after another with each new tree trained to reduce the remaining errors by addressing the current residual errors. The final prediction is made by adding up the outputs from all the trees.[2]

Gradient Boosting Regressor iteratively builds decision trees to minimize forecasting errors by focusing on residuals at each stage. In Indian air quality modeling, it helps capture intricate seasonal and temporal dependencies, such as the post-monsoon spike in PM2.5 or winter inversion effects.

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x)$$

Where:

- $f_m(x)$ = Prediction at iteration m,
- γ_m = Learning rate,
- $h_m(x)$ = New tree trained on previous residuals.

3.3. ADABOOST REGRESSOR

AdaBoost (short for Adaptive Boosting) is a statistical classification and regression meta-algorithm. It can be used in conjunction with many types of learning algorithms to improve performance. The output of multiple weak learners is combined into a weighted sum that represents the final output of the boosted classifier or regressor. Usually, AdaBoost is presented for binary classification, although it can be generalized to multiple classes or bounded intervals of real values.[3]

In the context of Indian air quality forecasting, this method helps focus on extreme pollution events or anomalies, such as sudden spikes in PM10 during Diwali or harvest burning seasons. By assigning higher importance to difficult-to-predict data points, AdaBoost enhances model responsiveness to critical air quality deterioration patterns.

$$f(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

Where:

- f(x) = Final AQ prediction,
- α_m = Weight for each weak learner,
- M = Total weak learners (e.g., decision stumps).
- $h_m(x)$ = New tree trained on previous residuals

3.4. HISTOGRAM GRADIENT BOOSTING REGRESSOR

Training the trees that are added to the ensemble can be dramatically accelerated by discretizing (binning) the continuous input variables to a few hundred unique values. Gradient boosting ensembles that implement this technique and tailor the training algorithm around input variables under this transform are referred to as histogram-based gradient boosting ensembles. It is common to refer to a gradient boosting algorithm supporting "histograms" in modern machine learning libraries as a histogram-based gradient boosting.[4]

$$f_m(x) = f_{m-1}(x) + \eta \cdot h_m(x)$$

Where:

- η = Learning rate,
  Histogram binning is applied to x before tree fitting,
- $f_m(x)$ = Model after m iterations.
- $h_m(x)$ = New tree trained on previous residuals.

3.5. XGBOOST REGRESSOR

XGBoost (eXtreme Gradient Boosting) is an open-source software library which provides a regularizing gradient boosting framework. XGBoost works as Newton–Raphson in function space unlike gradient boosting that works as gradient descent in function space, a second order Taylor approximation is used in the loss function to make the connection to Newton–Raphson method.[5]

$$f(x) = \sum_{k=1}^{K} h_k(x) + \Omega(h_k)$$

Where:

- $\Omega(h_k)$ = a regularization term,
- $h_k(x)$ = kth base model
- K = Number of boosting rounds.

4. DESIGN AND METHODOLOGY

The design of the proposed system for Indian Air Quality Forecasting using Ensemble Methods for Early Warning Systems follows a structured pipeline to ensure optimal performance and reliability. The process initiates with a comprehensive data preparation phase, where raw air quality datasets, including pollutants like PM2.5, NO2, SO2, and meteorological variables, are preprocessed. This involves handling missing values, scaling features, and transforming the data to ensure consistency and

robustness for downstream modeling tasks. Special attention is given to temporal patterns and regional variations in air quality levels across different Indian cities.

Once the data is cleaned and structured, it is fed into a suite of advanced ensemble-based regression models. These include the Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, Histogram-Based Gradient Boosting Regressor, and XGBoost Regressor. Each of these algorithms is designed to capture complex, non-linear dependencies in the atmospheric data, learning from patterns such as seasonal fluctuations, pollutant interactions, and meteorological influences to forecast future air quality indices (AQI) with high precision.

Following model training, a rigorous evaluation phase is conducted to compare the predictive capabilities of these ensemble methods. The models are assessed on unseen test data to ensure that the system generalizes well to real-world scenarios. The performance insights derived from this phase guide the next crucial step: hyperparameter tuning. Using techniques such as grid search or random search, the hyperparameters of each model are optimized to achieve the best possible forecasting accuracy while preventing overfitting.

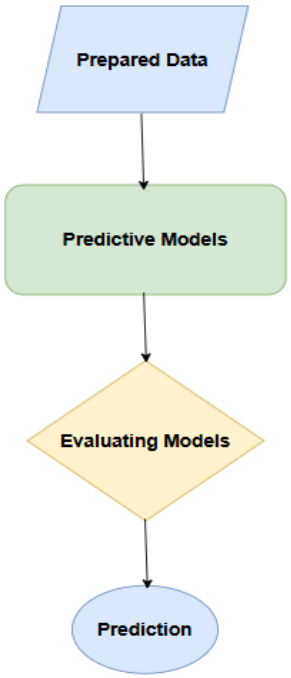

Fig -4: High Level Flow of Overall System

Finally, the optimized models are deployed to make predictions on unseen examples, serving as the core of the early warning system. These forecasts provide actionable insights for policymakers and the public, enabling timely interventions to mitigate the adverse effects of poor air quality.

5. RESULTS

Upon completing the training phase for all the ensemble models, each regressor was evaluated based on its predictive performance on the unseen test dataset. Among the models assessed—Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, Histogram Gradient Boosting Regressor, and XGBoost Regressor—XGBoost consistently emerged as the top-performing model. It achieved the lowest Root Mean Square Error (RMSE) on both the training and test sets, demonstrating superior accuracy and generalization capability.

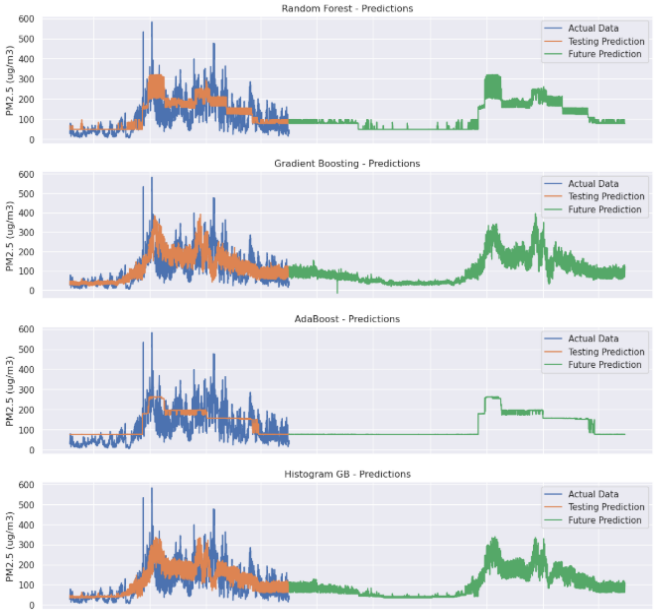

Fig -5(A): Prediction by other ensemble models

Based on these outcomes, XGBoost was selected as the final model for deployment within the air quality forecasting system. Its balance of speed and precision ensures timely and reliable forecasts, which are critical for early warning systems aimed at mitigating the risks associated with deteriorating air quality in Indian cities.

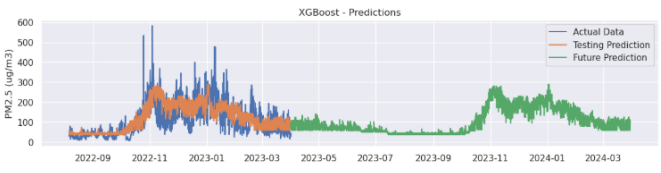

Fig -5(B): Prediction by XGBoost Model

6. CONCLUSIONS

This study focused on developing an effective early warning system for air quality forecasting in India using ensemble learning techniques. Various ensemble regressors, including Random Forest, Gradient Boosting, AdaBoost, Histogram Gradient Boosting, and XGBoost, were implemented and tested on prepared air quality datasets. After thorough evaluation, XGBoost emerged as the most reliable and efficient model, achieving the lowest RMSE values and outperforming others in both training and prediction speed. The project demonstrates that ensemble methods, particularly XGBoost, are highly capable of capturing intricate patterns in environmental data and producing timely forecasts. By deploying this model, the proposed system can serve as a valuable tool in predicting poor air quality events, enabling authorities to take proactive measures to protect public health. The findings emphasize the importance of model selection in designing accurate and scalable early warning systems for real-world environmental applications.

REFERENCES

[1] AnalytixLabs, (2023). Random Forest Regression — How it Helps in Predictive Analytics? on Medium. https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4

[2] Baladram, S. (2024). Gradient Boosting Regressor, Explained: A Visual Guide with Code Examples on Medium. https://medium.com/data-science/gradient-boosting-regressor-explained-a-visual-guide-with-code-examples-c098d1ae425c

[3] AdaBoost on Wikipedia. https://en.wikipedia.org/wiki/AdaBoost

[4] Brownlee, J. (2021). Histogram-Based Gradient Boosting Ensembles in Python on Machine Learning Mastery. https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/

[5] XGBoost on Wikipedia. https://en.wikipedia.org/wiki/XGBoost

[6] Singh, Kunwar P., et al. "Linear and nonlinear modeling approaches for urban air quality prediction." Science of the Total Environment (2012).

[7] Singh, Kunwar P., Shikha Gupta, and Premanjali Rai. "Identifying pollution sources and predicting urban air quality using ensemble learning methods." Atmospheric Environment (2013).

[8] Emeç, M. U. R. A. T., and M. Yurtsever. "A novel ensemble machine learning method for accurate air quality prediction." International Journal of Environmental Science and Technology (2025).

[9] Kumar, K., and B. P. Pande. "Air pollution prediction with machine learning: a case study of Indian cities." International Journal of Environmental Science and Technology 20.5 (2023): 5333-5348.

[10] Kumar, Anikender, and P. Goyal. "Forecasting of daily air quality index in Delhi." Science of the Total Environment 409.24 (2011): 5517-5523.

[11] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media, Inc.