International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Air Pollution Monitoring and Prediction using Machine Learning Algorithms

R. B. Dhumale[1], Ishani Cheke[2], Sanskruti Kakade[3], Srishti Salve[4]

[1,2,3,4]*Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, Maharashtra, India - 411001*

Abstract -

Air pollution is an increasing environmental concern worldwide, with hazardous impacts on the environment and human health. Monitoring and predicting air pollution levels accurately are crucial for implementing and mitigating its adverse effects. The purpose of this work is to forecast the value of the Air Quality Index using data preprocessing, feature extraction and selection, and machine learning based prediction techniques. The historical air quality data was collected and analysed from the dataset that included environmental data across multiple regions of India. Linear Regression, Decision Tree, Random Forest, XGBoost, RANSAC Regression, AdaBoost and LightGBM were used to process and forecast pollution concentrations. The Random Forest and LightGBM model had the highest prediction accuracy for Air Quality Index.

Key Words: Air pollution, environment and human health, Air Quality Index, Machine Learning, comparative analysis, ML models, prediction, historical data, India, Regression

1.INTRODUCTION

The inevitability of energy consumption and its repercussions is a reality in contemporary human activities. Air pollution comes from many human-induced sources, including industrial emissions from cars, airplanes, straw, burning coal and kerosene, and aerosol cans. Numerous dangerous pollutants, including CO, CO_2, PM, NO_2, SO_2, O_3, NH_3, and others, are released into the atmosphere on a daily basis. The health of people, animals, and plants is negatively impacted by the substances and particles that make up air pollution. This pollution can lead to severe human diseases, ranging from respiratory diseases such as tuberculosis, pneumonia, bronchitis, asthma, to fatal ailments such as lung cancer [1]. The World Health Organization (WHO) estimates that, relative to other causes of death, air pollution accounts for about 2.4 million deaths globally [2].

Air pollution not only negatively influences health but also has a significant and detrimental impact on the economy [1]. The study evaluated three machine learning methods: Support Vector Machine (SVM), Random Forest, and XG Boost, with XG Boost outperforming both SVM and Random Forest in accurately predicting air quality levels in a specific city [3]. Additionally, poor air conditions cause other environmental problems such as bad weather, global warming, rain or storm, reduced visibility, fog, aerosol formation, climate change and premature death [4].

The **Air Quality Index (AQI)** is a numerical rating that converts air quality data into an understandable value, providing an accurate depiction of local air quality. The significance of the AQI lies in its ability to convey crucial information about the quality of the air in a given location [5]. The motivation behind this project is twofold. Firstly, it directly impacts the well-being of communities by providing air quality data, enabling individuals to take informed decisions. Secondly, by predicting air quality, it offers authorities and environmental agencies a tool to implement pollution control measures effectively.

It was found that creating a statistical model that can predict such events within the scope of data analysis is quite difficult. It is a non-linear process with more impurities and they are not fully understood [6]. To address a potential issue of limited data scope, data used for model implementation should be collected from a wider range of areas [7].

M. Castelli introduced a proprietary machine learning fr amework designed to forecast air quality in California [8]. Using sensor data gathered from three different locations in Delhi, India's Capital City, and the National Capital Region (NCR), linear regression was used to study AQI prediction with data from Internet of Things setups [9]. Liang et al. investigated the performance of six machine learning classifiers in predicting Taiwan's AQI. It was concluded that Adaptive Boosting (AdaBoost) and Stacking Ensemble approaches are most suited for predicting air quality [10].

The proposed machine learning model, which combines Grey Wolf Optimization and Decision Tree, accurately

predicts AQI in major Indian cities, outperforming traditional algorithms. Experimentally verified, it achieves maximum accuracy of 88.98% in New Delhi, Bangalore, Kolkata, Hyderabad, Chennai, and Visakhapatnam [11]. The air quality index dataset utilized six machine-learning algorithms, with Random Forest and Decision Tree algorithms achieving the highest accuracy of 99.0%, outperforming each other[12]. The paper uses the Bayesian network model to predict air quality in Hangzhou, achieving an accuracy of over 80% and a close match to the actual value, indicating its practical value in air quality prediction[13].

The effective regulation of air quality relies on accurate predictions. However, amongst the researches conducted and work done till date, an evident pattern that emerges is the use of conservative algorithms only. Furthermore, future study suggests using sampling approaches such as the Synthetic Minority Oversampling Technique (SMOTE) to improve accuracy.

This paper utilizes rare algorithms like Light Gradient Boosting Machine (LightGBM) and Random Sample Consensus (RANSAC) as an alternative to traditional approaches that rely on intricate computations. The effect of SMOTE is also inspected as a valuable framework for predicting AQI.

2. METHODOLOGY

The proposed methodology shown in Fig. 1 provides a detailed account of the procedures and instruments utilized in this research. By means of methodical data collecting, extracting analysis, and comprehensive validation procedures, the technique was designed to meet the objectives of the research.

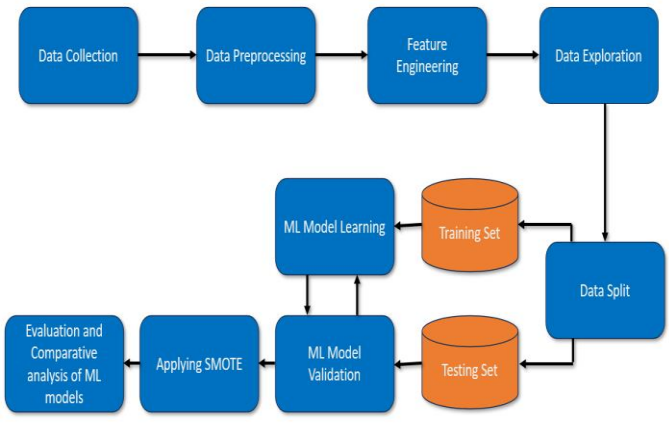

Fig. 1 Methodology to predict AQI using Machine Learning Models

The approach began with the data collecting phase, which involved identifying and gathering datasets related to air quality. This phase is critical since the performance of the machine learning model is strongly dependent on the quality and comprehensiveness of the obtained data. The collected data includes pollutant concentrations such as PM2.5, PM10, NO2, and CO. Following data collection, the data preprocessing phase entailed cleaning the data to remove any undesired elements and prepare it for further analysis. This includes dealing with missing values, reducing outliers, and maintaining data consistency. Data normalization was a key part of this step since it brought disparate data characteristics into a uniform scale of measurement, avoiding any feature from influencing the model. This was accomplished using min-max scaling.

The focus turned to feature engineering, which entailed developing new features to increase the performance of machine learning models by converting raw data into useful features. This better captures the underlying trends that influence air quality. The data exploration technique helped to analyse and visualize the data in order to grasp its structure, uncover patterns, and detect abnormalities. The dataset was then partitioned into two sets: training and testing, which were utilized for model learning and validation, respectively.

To mitigate potential bias caused by class imbalance in the data, Synthetic Minority Oversampling Technique (SMOTE) was applied. In order to ensure the development of a reliable and accurate model, the effectiveness of several machine learning algorithms in forecasting contaminants that affect air quality was assessed using statistical measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

3. IMPLEMENTATION

The data was gathered (across the years and states) from the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India (CPCB) under the National Data Sharing and Accessibility Policy (NDSAP). The data set included 435743 rows with 13 columns. The variables in the data set were particulate matter 2.5, particulate matter 10, nitric oxide, nitric dioxide, nitric x-oxide, ammonia, carbon monoxide, and sulphur dioxide. The duplicated values were dropped to improve the performance and generalization of the

model. The variables that were considered for the prediction of AQI were: Sulphur Index (si), Nitric Index (ni), Respirable Particulate Matter Index (rpi), Suspended Particulate Index (spi).

Variable Name	No. of Samples	Mean of values	Standard Deviation of values
si	435742.000000	12.361537	12.431079
ni	435742.000000	34.868097	29.950695
rpi	435742.000000	88.978039	45.499823
spi	435742.000000	82.317442	145.945641

Table 1. Statistics of pollutants that are considered in the prediction of AQI

The Table 1 provides an overall layout of the distribution of data in a descriptive manner, allowing a better understanding of the dataset. The number of samples were seen to be consistent throughout the variables. The mean and standard deviation of the four variables namely, si, ni, rpi and spi were also calculated.

The dataset was divided into three categories: 50% training data, 25% testing data, and 25% validation data. Regression models such as Linear Regression, Decision Tree, Random Forest, XGBoost, RANSAC Regression, AdaBoost, and LightGBM were used. Despite their computational demands, ensemble methods were deemed the most suitable for the study due to their superior accuracy in forecasting AQI trends for air pollution. By training regression models on past data, insights were gained for anticipating AQI levels.

SMOTE is an oversampling method designed specifically targets the minority class in imbalanced datasets. It accomplishes this by identifying k-nearest neighbors that represent existing data points within the minority class exhibiting similar characteristics. Subsequently, it interpolates between these neighbors to generate new synthetic data points. This process effectively augments the size of the minority class, leading to a more balanced dataset. SMOTE helps to ensure that the models are not biased towards the majority class and can learn more generalizable patterns from the data.

Evaluation metrics are quantitative measures used to analyse model performance. The most widely used metrics are Mean Average Error (MAE) as given in Eq 1 and Root Mean Square Error (RMSE) as given in Eq 2, which are calculated by comparing the anticipated and actual values. Furthermore, R- Sqared (R^2) as given in Eq 3 is an important metric for determining the strength of the association between predictive models and target variables.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{---Eq 1}$$

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

$$R^2 = 1 - \frac{\sum_i (yi - \hat{y}i)^2}{\sum_i (yi - \bar{y}i)^2} \quad \text{---Eq 3}$$

When compared in terms of the evaluation metrics, RandomForest and LightGBM stand out as the preferred option for predicting air quality in contrast to Linear Regression, Decision Tree, XGBoost, AdaBoost and RANSAC Regression. The strength of RandomForest lies in its capacity to integrate regularization techniques, ensuring high accuracy, scalability, efficiency, and flexibility. Despite Random Forest's superiority, there is recognition of certain areas where enhancements can be implemented. Addressing issues of overfitting, scalability and handling imbalanced data can further increase the performance of the model. Handling imbalanced data is recognized as a common issue across majority of the algorithms. This paper looks at solving this issue by using data handling techniques like SMOTE. It is applied on all the algorithms and its effect on the same is then compared to check for any enhancements.

4. RESULTS AND DISCUSSION

The average AQI over time in India is shown in Fig. 2 using a line graph. The x-axis shows the year, and the y-axis shows the AQI. The graph shows that the AQI has been increasing steadily over time, from around 60 in 1970 to over 160 in 2010.

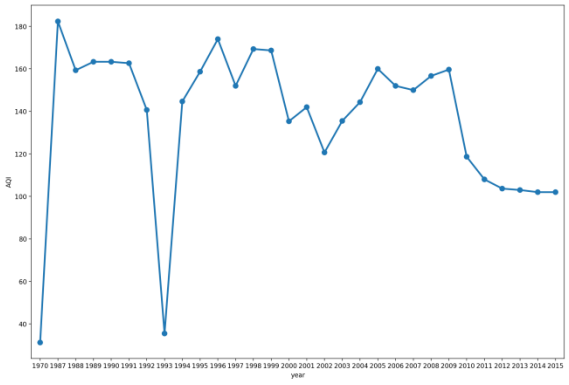

Fig. 2 Line graph showing average AQI over time in India

Following a comprehensive evaluation based on MAE, RMSE, R^2, and MAPE, two standout performers emerged: Random Forest and LightGBM. Random Forest exhibited superior performance owing to its ability to handle large dataset with high dimensionality and nonlinear relationships effectively. This quality was exceptionally advantageous as the dataset contained 435743 rows x 13 columns. Its ensemble learning approach, which aggregates the predictions of multiple decision trees, mitigates overfitting and enhances predictive accuracy. Moreover, its inherent feature selection capability proved advantageous in identifying the most influential factors contributing to AQI fluctuations. Similarly, LightGBM showcased remarkable performance by leveraging gradient boosting techniques to optimize model training speed and efficiency. Its ability to handle categorical features seamlessly, along with its efficient memory usage, rendered it particularly suitable for processing the diverse array of variables inherent in air quality prediction tasks. Additionally, LightGBM's ability to handle imbalanced datasets and its robustness against overfitting further solidified its position as an ideal choice for this predictive modelling. Ultimately, the combination of Random Forest and LightGBM not only delivered superior predictive performance but also useful insights into the complex dynamics of air quality variations in the goal state.

In order to enhance the predictive performance of the algorithms utilized in air quality prediction, SMOTE was implemented. It is a frequently used machine learning method for addressing class imbalance that generates synthetic samples from the minority class. By oversampling the minority class instances, the skewness in the dataset distribution can be corrected, preventing the model from being skewed in favour of the majority. In the context of predicting AQI, SMOTE proved to be advantageous by ensuring that the predictive models were trained on a more balanced representation of air quality scenarios, thus improving their ability to generalize to rare instances. Consequently, the predictive models refined with it exhibited enhanced robustness and accuracy, thereby strengthening their efficacy in forecasting AQI levels with greater precision and reliability. Overall, the integration of SMOTE into the modelling pipeline contributed significantly to refining the predictive capabilities of the algorithms, ultimately leading to more accurate and insightful air quality predictions for the targeted city.

Models	Without SMOTE				With SMOTE			
	MAE	RMSE	R^2	MSE	MAE	RMSE	R^2	MSE
Linear	0.959	1.132	0.220	1.282	1.197	1.308	0.413	1.711
Decision Tree	0.410	0.619	0.766	0.384	0.306	0.539	0.900	0.290
Random Forest	0.0010	0.038	0.9990	0.001	0.0016	0.049	0.9996	0.001
XGBoost	0.0184	0.143	0.987	0.020	0.018	0.133	0.993	0.017
RANSAC	0.910	1.436	-0.254	-	1.179	1.644	0.073	-
LightGBM	0.015	0.138	0.988	0.019	0.021	0.150	0.992	0.022
AdaBoost	0.622	0.706	0.696	0.741	0.529	0.7310.529	0.816	0.497

Table 2. Comparison of the prediction performance of selected models evaluated on four evaluation metrics.

The table summarizes the results of a comparative study on the effects of different data balancing techniques and machine learning models for a regression task. The models were evaluated primarily on two metrics: MAE and RMSE.

Among all the models tested, Random Forest and LightGBM achieved the best performance on both the imbalanced and balanced datasets. In the imbalanced dataset, Random Forest model had the lowest MAE which was 0.0010 and RMSE which was 0.038, while LightGBM had the second-lowest MAE which was 0.015 and RMSE which was 0.138. After using SMOTE, LightGBM model achieved the lowest MAE which was 0.018 and RMSE which was 0.133, while Random Forest model had the second-lowest MAE which was 0.0184 and RMSE which was 0.143.

The models were first trained on an imbalanced dataset, and then on a balanced dataset created using the SMOTE oversampling technique. The performance of all models improved after using SMOTE. For example, the Random Forest model's MAE decreased from 0.0010 to 0.0016, and its RMSE decreased from 0.038 to 0.049.

The results of this study suggest that Random Forest and LightGBM are both effective models for regression tasks, and that SMOTE can be a useful technique for improving the performance of models on imbalanced datasets.

5. CONCLUSION

In this study, an efficient model is created to forecast the value of the Air Quality Index using the air quality data provided by CPCB and NDSAP. Data preprocessing, feature extraction and selection, and machine learning based prediction techniques were performed on the dataset. In order to ensure an effective comparison, a comprehensive set of regression models was chosen that included Linear Regression, Decision Tree, Random Forest, XGBoost, RANSAC Regression, AdaBoost and LightGBM. The literature review conducted gave insights into the medical, economical and environmental concerns that come with air pollution. The machine learning lifecycle was used for prediction, splitting the data into training and testing sets. The performance for individual models was evaluated and compared using MAE, RMSE, R2 and MSE. SMOTE was applied to reduce possible bias resulting from class imbalance of the data. The findings reveal that Random Forest and LightGBM are effective models for regression tasks, and that SMOTE is a beneficial technique for improving model performance on unbalanced datasets.

6. REFERENCES

1. N. N. Maltare and S. Vahora, “Air Quality Index prediction using machine learning for Ahmedabad city,” *Digital Chemical Engineering*, vol. 7, Jun. 2023, doi: 10.1016/j.dche.2023.100093.

2. G. Mani, V. Joshi Kumar, and A. A. Stonier, “Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models,” *Journal of Engineering Research (Kuwait)*, vol. 10, no. 2 A, pp. 179–194, Jun. 2022, doi: 10.36909/jer.10253.

3. A. Deepak, A. S. Chavan, A. Bodhankar, L. Sherly Puspha Annabel, and A. Vanathi, “International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Advancing Air Quality Prediction in Specific Cities Using Machine Learning.” [Online]. Available: www.ijisae.org

4. K. Kumar and B. P. Pande, “Air pollution prediction with machine learning: a case study of Indian cities,” *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

5. V. Mani, “Predict the chennai AQI using Machine Learning and Time Series analysis.”

6. 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, Oct. 2023, doi: 10.1016/j.chemosphere.2023.139518.

7. X. Zhang, X. Jiang, and Y. Li, “Prediction of air quality index based on the SSA-BiLSTM-LightGBM model,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-32775-2.

8. M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8049504.

9. R. Kumar, P. Kumar, and Y. Kumar, “Time Series Data Prediction using IoT and Machine Learning Technique,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 373–381. doi: 10.1016/j.procs.2020.03.240.

10. Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, “Machine learning-based prediction of air quality,” *Applied Sciences (Switzerland)*, vol. 10, no. 24, pp. 1–17, Dec. 2020, doi: 10.3390/app10249151.

11. S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, “Optimized machine learning model for air quality index prediction in major cities in India,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-54807-1.

12. A. Pant, S. Sharma, and K. Pant, “Evaluation of Machine Learning Algorithms for Air Quality Index (AQI) Prediction,” *Journal of Reliability and Statistical Studies*, vol. 16, no. 2, pp. 229–242, 2023, doi: 10.13052/jrss0974-8024.1621.

13. Z. Fu, H. Lin, B. Huang, and J. Yao, “Research on air quality prediction method in Hangzhou based on machine learning,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Sep. 2021. doi: 10.1088/1742-6596/2010/1/012011.