IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 3, March 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203139 |

Harnessing Machine Learning for Enhanced Air Quality Prediction

Prof. Jaya Choubey[1], Prof. Divya Pandey[2], Prof. Mallika Dwivedi[3], Sneha Kushwaha[4], Riya Saraf[5]

Department of Computer Science Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India[1, 2, 3, 4, 5]

ABSTRACT: Air pollution poses significant health risks and necessitates the need for accurate air quality monitoring systems. This research paper focuses on predicting the Air Quality Index (AQI) using both regression and classification models. Various machine learning algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier were employed to estimate AQI based on pollutant concentrations of SO_2, NO_2, RSPM, and SPM. The performance of these models was evaluated using metrics like RMSE, R^2 score, accuracy, and Cohen's Kappa score. The results demonstrate the efficacy of the models in predicting AQI, with Random Forest models exhibiting superior performance in both regression and classification tasks. This study underscores the potential of machine learning techniques in enhancing the accuracy and reliability of air quality predictions, thereby contributing to better environmental management and public health policies.

KEYWORDS: Air Quality Index, Machine Learning, Regression Models, Classification Models, Environmental Monitoring, Predictive Analysis

I. INTRODUCTION

Air pollution has emerged as a critical environmental issue, adversely affecting the health and well-being of populations worldwide. The Air Quality Index (AQI) serves as a vital indicator for reporting daily air quality levels to the public, translating complex air quality data into an understandable and actionable format. Accurate prediction of AQI is essential for timely warnings and the implementation of necessary measures to mitigate the impact of poor air quality.

In recent years, machine learning techniques have gained prominence for their ability to analyze vast amounts of environmental data and provide reliable predictions. This study explores the use of various regression and classification algorithms to predict AQI based on the concentrations of key pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM).

The primary objectives of this research are to evaluate the performance of different machine learning models in predicting AQI, identify the most effective algorithms for this task, and analyze the implications of these findings for environmental monitoring and public health. By leveraging the strengths of both regression and classification approaches, this study aims to provide a comprehensive analysis of the predictive capabilities of machine learning techniques in the context of air quality assessment.

Researchers have recognized the detrimental effects of air pollution on historical monuments (Rogers 2019). Emissions from vehicles, power plants, factories, and agricultural activities contribute significantly to the increase in greenhouse gases, which negatively impact climate conditions and plant growth (Fahad et al. 2021a). These emissions also disrupt plant-soil interactions (Fahad et al. 2021b). Climatic changes affect not only humans and Quality Index (AQI) directly relates to public health, with higher AQI levels indicating more hazardous exposure for the population. This has driven scientists to monitor and model air quality, making the prediction of AQI, particularly in urban areas, a crucial and challenging task animals but also agricultural factors and productivity, leading to economic losses (Sönmez et al. 2021). The Air due to increasing vehicular and industrial activities. While most air quality studies focus on developing countries, where pollutants like PM2.5 are highly concentrated (Rybarczyk and Zalakeviciute 2021), only a few researchers have studied air quality prediction for Indian cities. The literature review highlighted a significant gap, prompting the need to analyze and predict AQI for India.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203139 |

Various models, including statistical, deterministic, physical, and Machine Learning (ML) models, have been used to predict AQI. Traditional probability and statistics-based techniques are often complex and less efficient, whereas ML-based AQI prediction models have proven to be more reliable and consistent. Advances in technology and sensors have facilitated precise data collection. The vast environmental data requires rigorous analysis, which ML algorithms can handle effectively. Al-Jamimi et al. (2018) discussed the significance of supervised ML algorithms for environmental protection issues. This study examines six years of air pollution data from Indian cities, analyzing twelve air pollutants and AQI. After preprocessing and cleaning the dataset, data visualization techniques are applied to uncover patterns and trends. This research leverages the correlation coefficient with ML models, a method used by few scholars (Alade et al. 2019a). Data imbalance is addressed using resampling techniques, and five popular ML models are evaluated using standard performance metrics. These metrics are widely used by scholars in the ML application researchers such as Ayturan et al. (2020), Alade et al. (2019b), Al-Jamimi et al. (2019), and Al-Jamimi and Saleh (2019).

II. LITERATURE REVIEW

Gopalakrishnan (2021) utilized Google's Street View data combined with machine learning to predict air quality at various locations in Oakland, California, particularly targeting areas with unavailable data. He developed a web application that forecasts air quality for any neighborhood in the city. Sanjeev (2021) analyzed a dataset containing pollutant concentrations and meteorological factors, concluding that the Random Forest (RF) classifier performed best due to its resistance to overfitting.

Castelli et al. (2020) applied the Support Vector Regression (SVR) machine learning algorithm to forecast air quality in California, focusing on pollutant and particulate levels. They introduced a novel method for modeling hourly atmospheric pollution. Doreswamy et al. (2020) investigated machine learning predictive models for forecasting particulate matter concentrations in the air, using six years of air quality monitoring data from Taiwan. They reported that their predicted values closely matched the actual values. Liang et al. (2020) evaluated the performance of six machine learning classifiers for predicting the AQI of Taiwan based on 11 years of data. They found that Adaptive Boosting (AdaBoost) and Stacking Ensemble were the most effective for air quality prediction, though performance varied by geographical region.

Madan et al. (2020) reviewed twenty studies on pollutants, machine learning algorithms, and their performance, finding that incorporating meteorological data like humidity, wind speed, and temperature improved pollution level predictions. Neural Networks (NN) and boosting models were found to outperform other machine learning algorithms. Madhuri et al. (2020) highlighted the significant impact of wind speed, wind direction, humidity, and temperature on air pollutant concentrations, employing supervised machine learning techniques to predict AQI and identifying the RF algorithm as having the fewest classification errors. Monisri et al. (2020) collected air pollution data from various sources and developed a mixed model to predict air quality, aiming to assist small towns in analyzing and forecasting air quality. Nahar et al. (2020) created a model using machine learning classifiers to predict AQI, based on 28 months of data from Jordan's Ministry of Environment, accurately identifying the most polluted areas. Patil et al. (2020) reviewed literature on various machine learning techniques for AQI modeling and forecasting, noting that Artificial Neural Networks (ANN), Linear Regression (LR), and Logistic Regression (LogR) were the most commonly used models for AQI prediction.

III. PROPOSED METHODOLOGY

III-A. Data Collection and Preprocessing

The dataset used in this study was obtained from a reputable air quality monitoring source, comprising 435,742 entries with 13 attributes including pollutant concentrations (SO_2, NO_2, RSPM, SPM) and various metadata. The preprocessing steps involved handling missing values, encoding categorical variables, and normalizing the data to ensure compatibility with the machine learning algorithms.

III-B. Calculation of Pollutant Indices

Individual pollutant indices for SO_2, NO_2, RSPM, and SPM were calculated using standard formulas. These indices were then used to compute the overall AQI for each entry in the dataset. The calculation methodologies for each pollutant index are as follows:

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203139 |

- SO_2 Index (SOi):

$$\text{SOi} = \begin{cases} \text{SO}_2 \times \left(\frac{50}{40}\right) & \text{if } \text{SO}_2 \leq 40 \\ 50 + (\text{SO}_2 - 40) \times \left(\frac{50}{40}\right) & \text{if } 40 < \text{SO}_2 \leq 80 \\ 100 + (\text{SO}_2 - 80) \times \left(\frac{100}{300}\right) & \text{if } 80 < \text{SO}_2 \leq 380 \\ \ldots & \text{and so on} \end{cases}$$

- NO_2 Index (Noi):

$$\text{Noi} = \begin{cases} \text{NO}_2 \times \left(\frac{50}{40}\right) & \text{if } \text{NO}_2 \leq 40 \\ 50 + (\text{NO}_2 - 40) \times \left(\frac{50}{40}\right) & \text{if } 40 < \text{NO}_2 \leq 80 \\ 100 + (\text{NO}_2 - 80) \times \left(\frac{100}{100}\right) & \text{if } 80 < \text{NO}_2 \leq 180 \\ \ldots & \text{and so on} \end{cases}$$

- RSPM Index (Rpi):

$$\text{Rpi} = \begin{cases} \text{RSPM} \times \left(\frac{50}{30}\right) & \text{if } \text{RSPM} \leq 30 \\ 50 + (\text{RSPM} - 30) \times \left(\frac{50}{30}\right) & \text{if } 30 < \text{RSPM} \leq 60 \\ 100 + (\text{RSPM} - 60) \times \left(\frac{100}{30}\right) & \text{if } 60 < \text{RSPM} \leq 90 \\ \ldots & \text{and so on} \end{cases}$$

- SPM Index (SPMi):

$$\text{SPMi} = \begin{cases} \text{SPM} \times \left(\frac{50}{50}\right) & \text{if } \text{SPM} \leq 50 \\ 50 + (\text{SPM} - 50) \times \left(\frac{50}{50}\right) & \text{if } 50 < \text{SPM} \leq 100 \\ 100 + (\text{SPM} - 100) \times \left(\frac{100}{150}\right) & \text{if } 100 < \text{SPM} \leq 250 \\ \ldots & \text{and so on} \end{cases}$$

III-C. Model Implementation

Multiple machine learning models were implemented to predict the AQI based on the calculated pollutant indices. The models included Linear Regression, Decision Tree Regressor, Random Forest Regressor for regression tasks, and Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors Classifier for classification tasks. The dataset was split into training and testing sets with an 80-20 ratio for regression models and a 67-33 ratio for classification models.

IV. MODEL EVALUATION AND RESULTS

The performance of the regression models was evaluated using Root Mean Squared Error (RMSE) and R-squared (R^2) scores. For classification models, accuracy and Cohen's Kappa score were used. These metrics provide a comprehensive understanding of the models' predictive capabilities.

IV-A. Model Training and Testing

- Linear Regression**:** Achieved high R^2 scores indicating strong predictive power.
- Decision Tree Regressor: Showed near-perfect fit on training data but slightly lower performance on test data.
- Random Forest Regressor**:** Provided the best performance with minimal RMSE and high R^2 scores on both training and test data.
- Logistic Regression**:** Demonstrated moderate accuracy and Kappa scores.
- Decision Tree Classifier**:** Achieved almost perfect accuracy and Kappa scores on both training and test data.
- Random Forest Classifier**:** Similar to Decision Tree Classifier, but with slightly higher robustness.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203139 |

- K-Nearest Neighbors Classifier**:** Showed high accuracy and Kappa scores, slightly lower than tree-based classifiers.

Model	RMSE (Train)	RMSE (Test)	R^2 (Train)	R^2 (Test)	Accuracy (Train)	Accuracy (Test)	Kappa Score (Test)
Linear Regression	13.58	13.67	0.9849	0.9847	-	-	-
Decision Tree Regressor	2.25e-13	1.30	1.0	0.9999	-	-	-
Random Forest Regressor	0.44	1.15	0.9999	0.9999	-	-	-
Logistic Regression	-	-	-	-	0.728	0.727	0.584
Decision Tree Classifier	-	-	-	-	1.0	0.9998	0.9997
Random Forest Classifier	-	-	-	-	1.0	0.9998	0.9998
K-Nearest Neighbors Classifier	-	-	-	-	0.998	0.997	0.995

Table I Comparison of different regression and classification parameters

V. CONCLUSION

This study demonstrates the efficacy of various machine learning models in predicting the Air Quality Index (AQI) based on pollutant concentrations. The Random Forest Regressor and Random Forest Classifier emerged as the top-performing models, showcasing exceptional predictive accuracy and robustness. These models provide valuable tools for environmental monitoring, enabling authorities to make informed decisions and take proactive measures to mitigate air pollution. The insights gained from this research can significantly contribute to public health and environmental management strategies.

REFERENCES

[1] Alade IO, Rahman MAA, Saleh TA (2019a) Predicting the specific heat capacity of alumina/ethylene glycol nanofluids using support vector regression model optimized with Bayesian algorithm. Sol Energy 183:74–82. https://doi.org/10.1016/j.solener.2019.02.060

[2] Alade IO, Rahman MAA, Saleh TA (2019b) Modeling and prediction of the specific heat capacity of $Al_2 O_3$/water nanofluids using hybrid genetic algorithm/support vector regression model. Nano-Struct Nano-Objects 17:103–111. https://doi.org/10.1016/j.nanoso.2018.12.001

[3] Al-Jamimi HA, Saleh TA (2019) Transparent predictive modelling of catalytic hydrodesulfurization using an interval type-2 fuzzy logic. J Clean Prod 231:1079–1088. https://doi.org/10.1016/j.jclepro.2019.05.224

[4] Al-Jamimi HA, Al-Azani S, Saleh TA (2018) Supervised machine learning techniques in the desulfurization of oil products for environmental protection: a review. Process Saf Environ Prot 120:57–71. https://doi.org/10.1016/j.psep.2018.08.021

[5] Al-Jamimi HA, Bagudu A, Saleh TA (2019) An intelligent approach for the modeling and experimental optimization of molecular hydrodesulfurization over AlMoCoBi catalyst. J MolLiq 278:376–384. https://doi.org/10.1016/j.molliq.2018.12.144

[6] Ayturan YA, Ayturan ZC, Altun HO, Kongoli C, Tuncez FD, Dursun S, Ozturk A (2020) Short-term prediction of PM2.5 pollution with deep learning methods. Global NEST J 22(1):126–131

[7] Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. https://doi.org/10.1186/s12889-017-4914-3

[8] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J ComputApplTechnol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203139 |

[9] Castelli M, Clemente FM, Popoviˇc A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/2020/8049504
[10] Dalberg (2019) Air pollution and its impact on business: the silent pandemic. https://www.cleanairfund.org/wp-content/uploads/2021/04/01042021_Business-Cost-of-Air-Pollution_Long-Form-Report.pdf
[11] Deshpande T (2021) India Has 9 Of World's 10 most-polluted cities, but few air quality monitors. indiaspend. https://www.indiaspend.com/pollution/india-has-9-of-worlds-10-most-polluted-cities-but-few-air-quality-monitors-792521
[12] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. ProcediaComputSci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221
[13] Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013
[14] Fahad, S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021b) Sustainable soil and land management and climate change (1st ed.). CRC Press. https://doi.org/10.1201/9781003108894
[15] Gopalakrishnan V (2021) Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71
[16] Gurjar BR (2021) Air pollution in india: major issues and challenges. energy future 9(2):12–27. https://www.magzter.com/stories/Education/Energy-Future/AIR-POLLUTION-IN-INDIA-MAJOR-ISSUES-AND-CHALLENGES
[17] IHME (2019) State of global air 2019 report. http://www.healthdata.org/news-release/state-global-air-2019-report
[18] Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning-based prediction of air quality. ApplSci 10(9151):1–17. https://doi.org/10.3390/app10249151
[19] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ICACCCN51052.2020.9362912
[20] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J SciTechnol Res 9(4):118–123
[21] Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G (2019) A machine learning model for air quality prediction for smart cities. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET). IEEE 452–457. https://doi.org/10.1109/WiSPNET45539.2019.9032734
[22] Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J AdvSciTechnol 29(5):6934–6943
[23] Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J AdvComputTechnol 9(9):3831–3840
[24] Patil RM, Dinde HT, Powar SK (2020) A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms 5(8):1148–1152
[25] Rogers CD (2019) Pollution's impact on historical monuments pollution's impact on historical monuments. SCIENCING. https://sciencing.com/about-6372037-pollution-s-impact-historical-monuments.html
[26] Rybarczyk Y, Zalakeviciute R (2017) Regression models to predict air pollution from affordable data collections. In: H. Farhadi (Ed.), Machine learning advanced techniques and emerging applications pp 15–48. IntechOpen. https://doi.org/10.5772/intechopen.71848
[27] Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys Res Lett. https://doi.org/10.1029/2020GL091202
[28] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538
[29] Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. https://doi.org/10.1201/9781003108931
[30] Soundari AG, Jeslin JG, Akshaya AC (2019) Indian air quality prediction and analysis using machine learning. Int J ApplEng Res 14(11):181–186

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details