Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526
Volume 14 Issue 01, 2025

Air Quality Index Prediction using Machine Learning

Aishwarya Patil[1]Pragati Patil[2]

[1]*Department of Computer Science, Shivaji University, patilaishwarya647@gmail.com*
[2]*Department of Computer Science, Shivaji University, Pragatipatil6571@gmail.com*

Peer Review Information

Submission: 17 Jan 2025
Revision: 14 Feb 2025
Acceptance: 15 March 2025

Keywords

Air Quality Index
Machine Learning
Environmental Data
Air Pollution
PM2.5
AQI Prediction

Abstract

Air pollution is a serious worldwide problem which negatively impacts on human health, climate, and ecosystems. Accurate AQI prediction is necessary for timely warning and environmental policy-making. In this paper, we investigate the use of different machine learning algorithms to forecast AQI based on environmental information like pollutant concentrations (PM2.5, PM10, NO_2, CO, SO_2, O_3), temperature, and humidity. The models are trained and tested on real-time and historical air quality datasets. This study compares the algorithms like Linear Regression, RandomForestRegressor, Support Vector Machine (SVM), XGBoostRegressor, GaussianNB. The outcomes reveal that ensemble-based models, specifically Linear Regression and XGBoostRegressor model offer high prediction accuracy.

Introduction

Air pollution may be described as a change in air quality that can be defined by the measurements of chemical, biological or physical contaminants in the air. Thus, air pollution refers to the unwanted presence of impurities or the abnormal increase in the ratio of some constituents of the atmosphere. It may be divided into 2 parts visible and invisible air pollution. Air pollution results from the occurrence in the atmosphere of poisonous elements, largely brought about by human activities, although at times it may be a result of natural events like volcanic eruptions, dust storms and forest fires, also stripping the air of quality. Air Quality Index (AQI) refers to a measure of air pollution in an area on a number scale. Air pollution is a significant worldwide health hazard, causing 7 million premature deaths every year (WHO). Pollution is rising due to urbanization, industrialization, vehicles, power plants, chemical processes, and some of the other natural processes like agricultural burning, volcanic eruptions, and wildfires. The reason for selecting this topic is air pollution impacts everyone's environment and health. Has AQI helps predict how dirty the air is, so we can take action Current air quality monitoring systems only report current conditions, but can't predict future air quality. We need a reliable way to forecast air quality levels, so we can take proactive steps to reduce pollution, protect public health, and create a sustainable future. AQI index between 0 and 500, utilized to convey how dirty the air is now or is predicted to be. The AQI incorporates five significant air pollutants which

include: particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), and sulfur dioxide (SO2). This project seeks to design a system capable of forecasting the quality of air we inhale. Through the application of machine learning, we can look at historical data and weather conditions to predict when air pollution will be high. This can be used by governments and citizens to take measures to minimize pollution and safeguard public health. Our vision is to create a tool that simplifies air quality prediction, makes it accurate, and makes it accessible to all.

Literature Review

Several studies have applied statistical and machine learning techniques for air pollution modeling and AQI forecasting.

Doreswamy et al. used various machine learning models to forecast air pollution in Taiwan, including linear regression, random forest, and gradient boosting regressor. The gradient boosting regressor model performed best, achieving an accuracy of 94.21% using data from 76 air pollution stations [1].

Chunhao Liu et al. proposed a Genetic Algorithm-based Improved Extreme Learning Machine (GA-IELM) and kernel extreme learning machine for Air Quality Index (AQI) forecasting. The model was tested on three real air quality datasets, covering pollutants like SO2, NO2, PM10, and CO, to improve the accuracy and reliability of air quality measurement [2].

Phuong N. Chau et al. used deep learning techniques to assess air quality in Quito, Ecuador during the COVID-19 lockdown, finding significant reductions in pollutant concentrations, such as CO (-48.75%), NO2 (-63.98%), and PM2.5 (-42.17%). The study used a Weather Normalized Models (WNM) approach, comparing the performance of five deep learning architectures to a Gradient Boosting Machine (GBM) model [3].

Mauro Casteli et al. used support vector regression (SVR) to analyze air quality data from California, USA, achieving an accuracy of 94.1% in predicting six AQI categories defined by the US EPA. The dataset, extracted from EPA's Air Quality, contained hourly data on pollutants like CO, SO2, NO2, and PM2.5, as well as temperature, humidity, and wind, with 102,090 records collected between 2016 and 2018[4].

Jingyang Wang et al. proposed a deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for predicting Air Quality Index (AQI), achieving high accuracy with MAE of 2.15, RMSE of 3.51, and R-squared of 0.92 [5].

K. Kumar and B.P. Pande used machine learning models to predict air pollution in Indian cities, with XGBoost performing best and showing high linearity between predicted and actual data. The dataset, sourced from India's Central Pollution Control Board (CPCB), covered major air pollutants like PM2.5, PM10, NO2, CO, SO2, and O3 from 2015 to 202 and was used to evaluate various models [6].

Rohit Kumar Singh et al. used machine learning algorithms to predict air quality, with Random Forest achieving the highest accuracy of 91.77%, outperforming Linear Regression and Gaussian Naïve Bayes [7].

Maltare and Vahora explored the effectiveness of machine learning models like SARIMA, SVM, and LSTM in predicting air quality index (AQI) in Ahmedabad City, evaluating their performance using metrics like coefficient of determination and root mean squared error [8].

Hieu Dao et al. used Random Forest, XGBoost, and Neural Network machine learning models to predict air quality, evaluating their performance using score error, RMSE, and coefficient of determination. The results showed that Random Forest and XGBoost had similar performance, while Neural Network was less efficient, with XGBoost emerging as the most effective model [9].

Ravindiran et al. used machine learning models like Random Forest and Catboost to accurately predict air quality in Visakhapatnam, India, achieving high correlations of 0.9998 and 0.9936, respectively [10].

N. Srinivasa Gupta et al. used machine learning models like SVR, RFR, and CatBoost regression to predict AQI in Indian cities, achieving a maximum accuracy of 97.6% in Kolkata using RFR [11].

Madhuri VM et al. used machine learning models to predict air quality, with Random Forest achieving the highest accuracy of 91.45% using a dataset from India's Central Pollution Control Board (CPCB) [12].

Methodology

The machine learning method for forecasting Air Quality Index (AQI) entails a few important steps that begin with the retrieval of proper environmental and meteorological data like pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, O_3), temperature, humidity, wind speed, and so on. After they are gathered, the data is preprocessed to deal with missing data, delete noise, normalize features, and obtain helpful patterns. If AQI values are not available directly,

they can be estimated based on standard pollutant concentration breakpoints specified by environmental agencies. Next, Exploratory Data Analysis (EDA) is conducted to learn about trends, correlations, and feature importance. Relevant features are then chosen based on the insights to train machine learning models. Based on the task, regression models (such as Random Forest Regressor, XGBoost, or LSTM) are applied to forecast continuous AQI values, whereas classification models (e.g., Decision Trees, SVM, or neural networks) are utilized to forecast categorical AQI levels. The chosen model is trained and cross-validated using methods such as cross-validation and optimized for maximum performance. Performance metrics like RMSE, MAE, R^2 score for regression, or accuracy, precision, recall, and F1-score for classification are employed to measure model performance. Lastly, the model is implemented for real-time or scheduled predictions and kept under continuous surveillance to ascertain accuracy, with regular updates in terms of new data to ensure durability over time.

Machine Learning Algorithms

Linear Regression

Linear regression is a supervised machine learning model employed for making predictions of continuous values such as the Air Quality Index (AQI). Linear regression tries to establish a linear relationship between input features like PM2.5, PM10, NO2, etc., and the target, AQI. It learns from historical data by minimizing the difference between the actual and predicted AQI using the least squares method. Once trained, the model will be able to predict AQI from new environmental information. While it is easy to implement and convenient to utilize, linear regression cannot possibly represent sophisticated patterns in air quality data but works well as a baseline model.

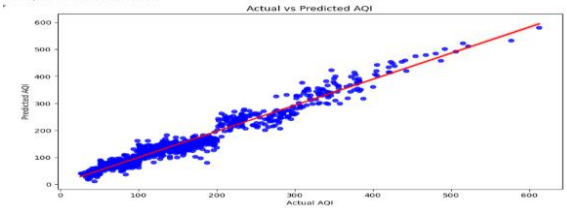

Fig. 1 Linear Regression

Random Forest

Random Forest is a machine learning ensemble algorithm that performs very well in predicting continuous values such as the Air Quality Index (AQI). It achieves this by training multiple decision trees and then averaging their predictions to enhance accuracy and minimize overfitting. In predicting AQI, inputs such as PM2.5, PM10, NO2, CO, and other environmental factors are utilized to train the model. Each tree in the forest is trained on a random subset of the data, which allows it to capture complicated, non-linear relationships within air quality patterns. Random Forest is stable, capable of dealing with missing data, and tends to produce high prediction accuracy and is therefore a reliable option for AQI prediction.

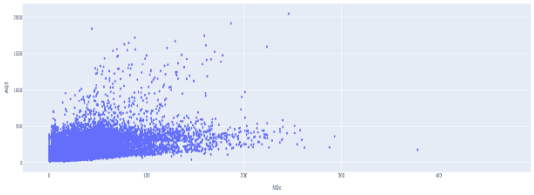

Fig. 2 Random Forest

Support Vector Regression (SVR)

Support Vector Machine regressor is an efficient supervised learning algorithm for making predictions of continuous values such as the Air Quality Index (AQI). SVM regressor searches for a hyperplane in the high-dimensional feature space that minimizes the error while keeping the error within some threshold (epsilon) and the margin between actual and predicted values as large as possible. In AQI forecasting, input variables like PM2.5, PM10, NO2, and other contaminants are utilized to train the model to learn patterns and trends in air quality. The SVM regressor can identify non-linear relationships with kernel functions and is thus ideal for intricate environmental data. Though it needs more computational power than linear regression, it tends to offer greater accuracy for AQI forecasting when appropriately tuned.

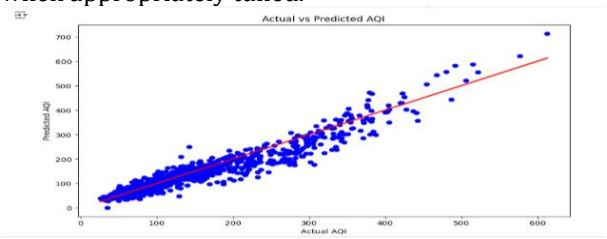

Fig. 3 Support Vector Regression (SVR)

XGBoost

XGBoost (Extreme Gradient Boosting) is a fast and effective machine learning algorithm commonly employed for the prediction of continuous values such as the Air Quality Index (AQI). It performs by creating an ensemble of decision trees sequentially, with each subsequent tree aimed at

correcting the mistakes made by the preceding ones, optimizing the model through gradient descent. In predicting AQI, XGBoost employs characteristics such as PM2.5, PM10, NO2, CO, and weather to learn intricate patterns in the data. XGBoost manages missing data, overfitting, and big data effectively using methods such as regularization and parallel computing. Because of its efficiency in accuracy and speed, XGBoost is most commonly used for accurate and consistent AQI forecasting.

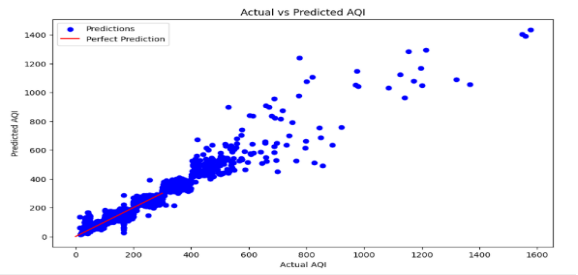

Fig. 4 XGBoost

GaussianNB

Gaussian Naive Bayes (Gaussian NB) is a probabilistic machine learning model that is generally employed for classification, but can be modified for regression-type tasks like Air Quality Index (AQI) prediction by classifying AQI into various levels (e.g., good, moderate, unhealthy). It operates on the basis of Bayes' Theorem under the assumption that the input features—like PM2.5, PM10, NO2, and CO—are normally distributed and independent of one another. The algorithm computes the probability of every AQI category using the input features and predicts the most probable category. Gaussian NB is efficient, easy, and does well even with small data when the independence assumption is marginally broken. It is less efficient for predicting the exact AQI value but can be beneficial in classifying air quality into health-related classes.

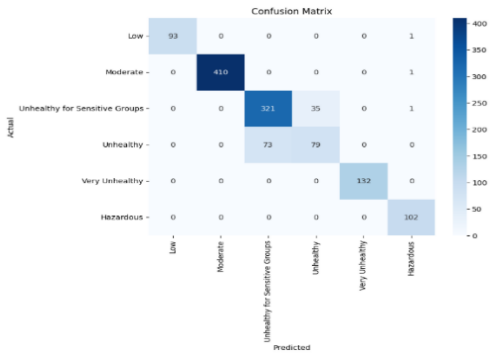

Fig. 5 GaussianNB

Result

The output of machine learning (ML) based air quality index (AQI) prediction normally entails the forecasting of concentrations of different air pollutants like particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). Prediction is carried out using models trained with past air quality records comprising environmental variables like temperature, humidity, wind speed, and location.

For instance, a machine learning model such as a Random Forest, Support Vector Machine (SVM), or deep learning model may be trained to forecast AQI values. The model takes input features such as time of day, meteorological data, and past levels of pollutants to project future AQI levels. The result is usually a forecasted AQI value, usually categorized into levels like "Good," "Moderate," "Unhealthy," etc., to guide people on the air quality of a specific location. The model's performance is measured based on metrics such as Mean Squared Error (MSE) or R-squared, depending on whether it is a regression or classification problem. By making precise AQI predictions, these machine learning algorithms assist with decision-making on public health, environmental policy, and air quality control.

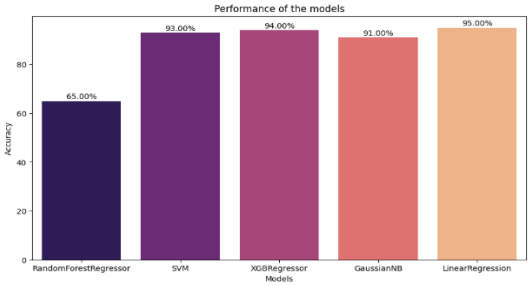

Fig.6 Comparision of all Algorithms

Future Scope

The future potential of air quality index (AQI) forecasting with machine learning is bright, with increasing capabilities in data acquisition through sensors and IoT devices having a positive effect on the accuracy of models. Hyper-local and real-time predictions will improve with time, and interventions in hotspot regions can be done promptly. Machine learning algorithms may incorporate more heterogeneous data, including socio-economic and behavioral factors, to deliver rich information on pollution trends. Moreover, AQI forecasting could be a critical component of smart city technology, allowing for the best traffic and industrial operations in real-time, based on

pollution predictions. With advancements in AI methods, the use of explainable AI (XAI) might enhance transparency, boosting confidence in the forecasts and supporting better public health and environmental policy.

Conclusion

In summary, machine learning has vast potential in air quality index (AQI) prediction, making more precise and timely predictions that can be used to reduce the effects of air pollution. ML models, using historical and real-time data, allow for better insight into patterns of pollution and their impact on public health. Predictions inform decision-making in urban planning, environmental policy, and public health interventions. As technology continues to evolve, machine learning algorithms will become finer and provide local and more accurate AQI forecasts. Overall, using machine learning for AQI forecasting will be an important enabler in designing smarter, healthier cities and enhanced environmental sustainability.

References

Doreswamy, Harishkumar K S, Yogesh KM, and Ibrahim Gad. "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models." Third International Conference on Computing and Network Communications (CoCoNet'19) 171 (January 1, 2020): 2057–66. https://doi.org/10.1016/j.procs.2020.04.221.

Liu, Chunhao, Guangyuan Pan, Dongming Song, and Hao Wei. "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine." IEEE Access 11 (2023): 67086 97. https://doi.org/10.1109/ACCESS.2023.3291146.

Chau, Phuong N., Rasa Zalakeviciute, Ilias Thomas, and Yves Rybarczyk. "Deep Learning Approach for Assessing Air Quality During COVID-19 Lockdown in Quito." Frontiers in Big Data 5 (April 4, 2022): 842455. https://doi.org/10.3389/fdata.2022.842455.

Castelli, Mauro, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. "A Machine Learning Approach to Predict Air Quality in California." Complexity 2020 (August 4, 2020): 1–23. https://doi.org/10.1155/2020/8049504.

Wang, Jingyang, Xiaolei Li, Lukai Jin, Jiazheng Li, Qiuhong Sun, and Haiyao Wang. "An Air Quality Index Prediction Model Based on CNN-ILSTM." Scientific Reports 12, no. 1 (May 19, 2022): 8373. https://doi.org/10.1038/s41598-022-12355-6.

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, no. 5 (May 2023): 5333–48. https://doi.org/10.1007/s13762-022-04241-5.

Kumar Singh, Rohit, Shekhar Raghav, Tarun Maini, Murari Kumar Singh, and Md. Arquam. "Air Quality Prediction Using Machine Learning." SSRN Electronic Journal, 2022. https://doi.org/10.2139/ssrn.4157651.

Maltare, Nilesh N., and Safvan Vahora. "Air Quality Index Prediction Using Machine Learning for Ahmedabad City." Digital Chemical Engineering 7 (June 1, 2023): 100093. https://doi.org/10.1016/j.dche.2023.100093.

Dao, To-Hieu, Hoang Van Nhat, Hoang Quang Trung, Vu Hoang Dieu, Nguyen Thi Thu, Duc-Nghia Tran, and Duc-Tan Tran. "Analysis and Prediction for Air Quality Using Various Machine Learning Models," 89–94, 2022. https://doi.org/10.15439/2022R03.

Ravindiran, Gokulan, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, and Christian Sonne. "Air Quality Prediction by Machine Learning Models: A Predictive Study on the Indian Coastal City of Visakhapatnam." Chemosphere 338 (October 1, 2023): 139518. https://doi.org/10.1016/j.chemosphere.2023.139518.

Gupta, N. Srinivasa, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, and G. Arulkumaran. "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis." Edited by Rahil Changotra. Journal of Environmental and Public Health 2023 (January 30, 2023): 1–26. https://doi.org/10.1155/2023/4916267.

Natarajan, Suresh Kumar, Prakash Shanmurthy, Daniel Arockiam, Balamurugan Balusamy, and Shitharth Selvarajan. "Optimized Machine Learning Model for Air Quality Index Prediction in Major Cities in India." Scientific Reports 14, no. 1 (March 21, 2024): 6795. https://doi.org/10.1038/s41598-024-54807-1.