SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

An Effective Air Pollution Prediction Model Using Machine Learning Algorithms

Kayalvizhi Subramanian[1,*], Gunasekar Thangarasu[2]

[1] Faculty of Engineering, Built Environment and Information Technology, MAHSA University, 42610, Saujana Putra, Selangor, Malaysia
[2] Department of Professional Industry Driven Education, MAHSA University, 42610, Saujana Putra, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 6 June 2023
Received in revised form 2 January 2024
Accepted 4 March 2024
Available online 25 June 2024

Keywords:

Machine Learning; Air pollution prediction; Linear regression; Artificial neural network and KNN

ABSTRACT

Air pollution is a major environmental concern globally, with both developed and developing countries facing its impacts. In recent years, citizens and governments have become increasingly concerned about the effects of air pollution on human health and have proposed sustainable development initiatives to address this issue. In a recent study, air pollution data from the years 2020-2022 was collected from a secondary data source. The data set included six key input features, including SO2, PM2.5, CO, PM10, NO2, and O3 values. To analyse this data, various machine learning models were employed, including linear regression, multiple linear regression, KNN, random forest regression, decision tree regression, support vector regression, and artificial neural networks. To ensure the accuracy of the predictions, mean square error and R square were used to measure the absolute error and forecast precision. Additionally, the importance of each input feature in air pollution was investigated, providing valuable insights into the factors that contribute to air pollution levels. Overall, the use of machine learning algorithms in air pollution estimation and prediction has significant potential to improve our understanding of this critical environmental issue and inform effective strategies for addressing it.

1. Introduction

Air pollution is a serious global issue that has a significant impact on both people and the environment. It occurs when harmful substances are present in the air, leading to negative effects on human health and the natural world. These substances can include various pollutants, such as particulate matter, nitrogen oxides, sulfur dioxide, carbon monoxide, and volatile organic compounds. Air pollution has numerous harmful effects, including respiratory illnesses, heart disease, and stroke, among others. It also harms the environment, contributing to climate change and damaging ecosystems. As such, addressing air pollution is a critical concern for governments, organizations, and individuals alike.

* *Corresponding author.*
E-mail address: skayalvizhi2012@gmail.com

https://doi.org/10.37934/araset.47.2.6875

To mitigate the effects of air pollution, various strategies can be employed, such as reducing emissions from vehicles and industrial sources, increasing the use of clean energy sources, and promoting sustainable transportation options. Additionally, monitoring air pollution levels and understanding their sources and impact is essential for effective policymaking and decision-making. Air pollution is caused by a variety of human activities, such as burning fossil fuels, industrial processes, agriculture and transportation. It is a global problem, affecting not only urban areas but also rural regions. In some cities, air pollution levels have reached dangerous levels, leading to increased rates of respiratory diseases, heart problems, and even premature death [1].

In addition to its impacts on human health, air pollution can also have negative effects on the natural environment, including acid rain, which can damage crops and forests and contribute to global warming [2]. The global community has taken steps to address this issue, such as signing international treaties and agreements, such as the Paris Agreement, which aims to reduce greenhouse gas emissions and limit the impacts of climate change. However, much work still needs to be done to reduce air pollution and protect the health of individuals and the planet as a whole. This requires a concerted effort from governments, industries and individuals to reduce emission, switch to cleaner energy sources and adopt more sustainable practices.

The significance of forecasting air pollution is a vital tool that enables proactive measures to safeguard public health, preserve the environment, and foster sustainable development. By anticipating and addressing potential pollution challenges, societies can work towards creating cleaner, healthier, and more resilient communities.

1.1 Objectives of the Study

Machine learning algorithms can be used to build predictive models that estimate the air pollution level based on various inputs such as weather conditions, emissions data and traffic patterns. These models can provide real-time air quality forecasts, which can help policymakers and individuals make informed decisions.

1.2 Forecasting Air Pollution Using Conventional Methods can face Several Challenges

Data quality and availability: Inaccurate or missing data can significantly impact the accuracy of the forecast. Data quality issues can arise due to measurement errors, incorrect readings, or missing data. Non-linear relationship: Air pollution is influenced by various factors such as weather patterns, traffic, industrial activities, etc., which can have non-linear relationships with each other. Conventional statistical methods may struggle or capture these complex relationships, leading to inaccurate in the forecast. Temporal and Spatial variability: Air pollution levels can vary significantly from one location to another and from one-time period to another. Forecasting models must take into account this temporal and spatial variability to produce accurate forecasts. Model limitations: Conventional statistical methods such as regression analysis or time series models have limitations in terms of their ability to handle complex relationships and large amounts of data, inaccuracies in the forecast can arise if the model is not properly configured or if it is not a good fit for the data.

2. Literature Review

The deterioration of air quality over the past six years has had a negative impact on citizens' quality of life. A recent study [3] investigated the effect of air pollution from vehicles on human health. Improving air quality forecasting is a critical goal for society. Major air pollutants include sulfur

dioxide, PM2.5, and nitrogen oxides. Sulfur dioxide, a gas present in the atmosphere, readily combines with other substances to form harmful substances such as sulfuric and sulfurous acids. When inhaled, sulfur dioxide can cause coughing, wheezing, shortness of breath, and tightness in the chest, as well as burning sensations in the nose, throat, and airways. The concentration of sulfur dioxide in the environment can also affect habitat suitability [4]. This study [5] focused seven regional monitoring areas in geographical and meteorological divergence for forecasting the air pollution results. Besides, they have compared the obtained forecast results with Taiwan, Taipei and London country outputs. Then, investigate the impact of industrial pollution and recommend the improved version of the prediction model to improve their prediction accuracy.

PM2.5, also known as fine particulate matter, is a major health concern when levels in the air are high. A study [6] examined seven regional monitoring areas with diverse geographical and meteorological conditions for air pollution forecasting. The results were compared to outputs from Taiwan, Taipei, and London. The study also analyzed the impact of industrial pollution and proposed improvements to the prediction model to enhance accuracy.

The researchers [7-8] used the widely popular machine learning technique of support vector regression (SVR) to forecast pollutant and particle levels and the Air Quality Index (AQI). Among the options considered, the radial basis function (RBF) kernel produced the most accurate predictions for SVR. The use of all available variables, rather than just a subset chosen through principal component analysis, was found to be a more effective method. SVR with an RBF kernel accurately predicted hourly pollutant concentrations for substances such as carbon monoxide, sulfur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5 and PM10.

A different study [9-10] evaluated the performance of predictive machine learning models for forecasting particulate matter concentration in the air using air quality monitoring data from Taiwan from 2012 to 2017. Mean absolute error (MAE), mean square error (MSE), and coefficient of determination were used as evaluation metrics. Machine learning techniques, such as artificial neural networks (ANNs), are widely used for air quality forecasting due to their ability to take into account a broad range of parameters [11]. Several studies have demonstrated the use of composite models or separate models based on regression algorithms. Other artificial intelligence algorithms include fuzzy logic, generative algorithms, and principal component analysis.

3. Methodology

The artificial intelligence techniques have an impact on our daily lives, and artificial intelligence-based algorithms are commonly used for prediction purposes, particularly for air quality forecasting. The process of using machine learning algorithms for air quality index (AQI) prediction typically involves several stages [12-15].

Data Collection: This involves collecting air quality data such as levels of various pollutants (e.g.. PM2.5, CO, NO2, etc.), meteorological parameters (e.g. Temperature, humidity, wind speed, etc.) And other relevant information that can affect air quality. Data Pre-processing: The collected data may need to be pre-processed to handle missing values, remove outliers, and perform the normalization to ensure the data is ready for analysis. Feature Engineering: This involves selecting and transforming relevant features of the collected data that can have a significant impact on air quality. Model Selection: This involves choosing a suitable machine learning model that can be used to predict AQI. Some commonly used models include decision trees, random forests, neural networks, etc. Model Training: This involves training the selected machine learning model of the pre-processed and engineering data. This stage typically involves optimizing the model's type parameters to obtain the best performance. Model Evaluation: This involves evaluating the performance of the

trained model on a test set, which is a set of data that was not used during the training stage. Common evaluation metrics for AQI prediction include mean absolute error, mean squared error, R-squared, etc. Model Deployment: Finally, the trained and evaluated model can be deployed in real-world applications, such as predicting AQI in a given location, to help individuals and organizations make informed decisions to mitigate the effects of air pollution. These stages can be iterated multiple times to further improve the model's performance, by fine tuning the features, type-parameters or by selecting a different machine learning model. The whole process steps of air quality forecasting show in Figure 1.

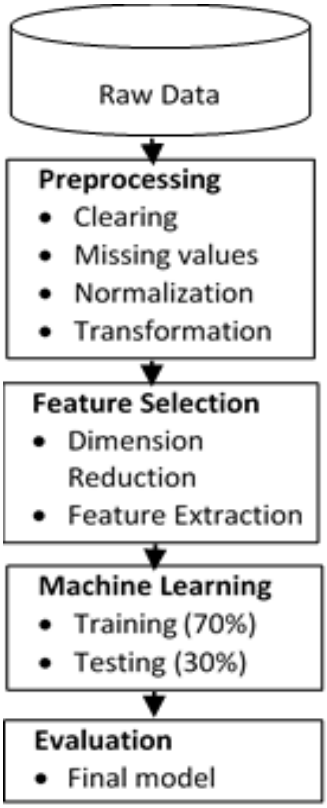

Fig. 1. Machine learning processing steps

3.1 Data Source

This study collected pollutant concentrations of all elements present in the air to forecast air quality in the National Capital Region (NCR) area covered with 55,083 kilometers around Delhi [11]. The data will be available on the Central Pollution Control Board for the year 2020-2022. We used data from several stations that measure a variety of atmospheric elements. The AQI formulae have been used to calculate the AQI for a specific year using various regression algorithms.

3.2 Air Quality Index Table

An air quality index [16] is used by government agencies to inform the public about how polluted the air is now and how polluted it will become in the future. The risks to public health grow as the AQI rises shown in the Table 1.

Table 1
Air Quality Index Table in India

AQI Colour	Level of Concern	Value of Index
Green	Good	0-50
Yellow	Moderate	51-100
Orange	Unhealthy for sensitive group	101-150
Red	Unhealthy	151-200
Purple	Very Unhealthy	201-300
Maroon	Hazardous	301 and Higher

Air quality indices differ from country and correspond to different national air quality standards. This study concentrated on the formula used in India to calculate the AQI.

3.3 Computing AQI

The air quality index is a numerical value used to describe the air quality in a given region. It provides a simple and easy-to-understand measure of air pollution levels, taking into account various pollutants in Eq. (1) [17].

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low} \quad (1)$$

where:
C_{low} = The concentration breaking point that is ≤ C
C_{high}= The concentration breaking point that is ≥ C
I_{low} = The index breaking point corresponding to C_{low}
I_{high} = The index breaking point corresponding to C_{high}

The AQI is calculated as follows.:
• Determine the concentration of each pollutant in the air. The can be measured by air quality monitoring stations or through remote sensing methods.
• Calculate the pollutant's sub-index for each pollutant by dividing the pollutant concentration by its breakpoint value.
• The AQI is the maximum sub-index value among the pollutants considered.
• Finally, the AQI is categorized into different levels of air quality based on a range of AQI values, with higher AQI values indicating poorer air quality.

It is important to note that the AQI calculation method may vary slightly depending on the country or region. However, the basic principle remains the same to provide a simple and easy-to-understand measure of air pollution levels.

3.4 Data Pre-processing

Data preprocessing [18-19] is a crucial stage in the machine learning process that ensures that the data being used for modeling is in the appropriate format, clean, and consistent. The objective of data preprocessing is to convert raw data into a format that is appropriate for machine learning models.

The common steps involved in data preprocessing in machine learning include [20]:

Data cleaning: The process of data cleaning includes identifying and correcting any discrepancies or omissions in the data. This may involve eliminating invalid or missing values, or correcting errors in the data set to ensure that it is accurate and complete.

Data transformation: Data transformation is a crucial step that involves altering the data to make it more appropriate for use in machine learning models. This can include several techniques, such as normalizing the data, converting categorical variables to numerical ones, or scaling the data to a particular range. The aim of this step is to create a more manageable and interpretable data set that can be used to build robust and accurate machine learning models.

Data reduction: Data reduction is the process of reducing the number of variables or features in the data set. This can be done to eliminate irrelevant or redundant information that may hinder model performance, or to simplify the data and make it easier to handle. Data reduction techniques can include feature selection or feature extraction methods, such as principal component analysis (PCA), which can reduce the dimensionality of the data set while retaining important information. The goal of data reduction is to create a more manageable and streamlined data set that is better suited for machine learning modelling.

Data splitting: Data splitting is a crucial step in machine learning that involves dividing the data set into three distinct subsets: the training set, the validation set, and the testing set. The training set is used to train the model, while the validation set is used to tune the model's hyper parameters and optimize its performance. Finally, the testing set is used to evaluate the model's overall performance and accuracy. The goal of data splitting is to ensure that the model is able to generalize well and perform accurately on unseen data by assessing its performance on a separate data set.

3.5 Training the Model

Data was split 70% for training and 30% for testing. There are numerous models in machine learning for predicting outcomes. Some are in favour of regression, whereas others are in favor of classification. Because we know that AQI has a distinct value, we must train the model using a regression algorithm. The researchers decided to use various algorithms such as support vector regression, multiple linear regression, Lasso-Ridge regression, and k-nearest neighbour, among others.

3.6 Model Evaluation

Evaluating the performance of a machine learning algorithm for forecasting air pollution index (AQI) is an important step to ensure that the model provides accurate and reliable predictions.

4. Result and Discussion

One of the most important goals of the research paper was to develop a machine learning model to predict air pollution. We predict air pollution based on specific data for the year, 2023. We use data from previous years obtained from Central Pollution Control Board website. Through the use of linear regression, Lasso-Ridge regression, KNN, and support vector machine.

Table 2 displays the mean absolute error and root mean squared error for the various machine learning algorithms utilized in the current study. These metrics are valuable for comparing and evaluating different algorithms, and can aid in the selection of the most effective model for a given problem.

Table 2
Root Mean and R Squared Values of Algorithms

Models	Means Absolute Error	Root Mean Squared Error
Multiple Linear Regression	39.6776	67.97359
Lasso Regression	39.6587	67.09765
Ridge Regression	39.3942	67.98634
Support Vector Regression	0.09756	0.392473
K-Nearest Neighbour	3.45643	21.86422

Following the evaluation of various types of regression models. The best fit model for predicting AQI is the Support Vector Regression model, which has a 97.5 percent accuracy. This model is the best match. Figure 2 depicts a comparison of actual and predicted values. Here, the study shows that almost all values are equal, but some values exhibit anomalous behaviour. If we look at the average of AQI month by month in Figure 3, we can easily conclude that AQI is affected by factors other than particle concentration. It concludes that we require more data and add more columns to the dataset of other factors such as temperature and so on.

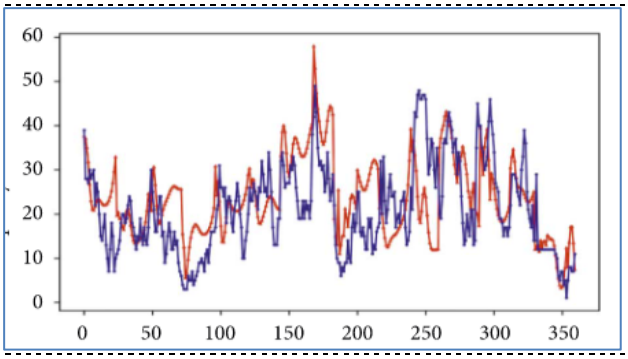

Fig. 2. Comparison between actual and predicted values

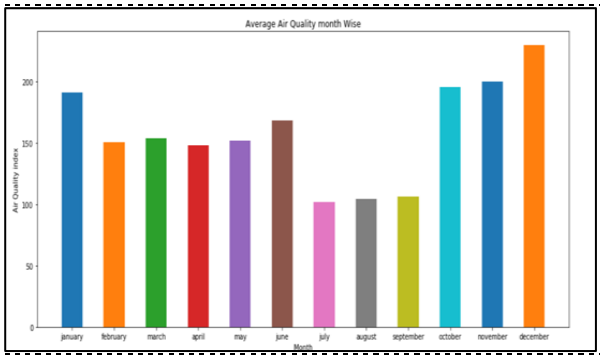

Fig. 3. The average forecasting result of AQI

Here can clearly see that the average obtained result for the year 2023 significantly differs from the same months in previous years.

5. Conclusion

In a recent extensive investigation, data on air pollution spanning from 2020 to 2022 was systematically collected from a dependable secondary data source. The dataset included crucial input features like SO2, PM2.5, CO, PM10, NO2, and O3 values. To derive meaningful insights from this extensive dataset, a variety of machine learning models were strategically utilized. These models covered linear regression, multiple linear regression, K-nearest neighbours (KNN), random forest regression, decision tree regression, support vector regression (SVR), and artificial neural networks. This specific analysis aimed to forecasting individual pollutant levels. The results indicated that the SVR method displayed significant effectiveness, producing a robust model for air pollution. This model exhibited admirable precision in precisely modelling concentrations of crucial pollutants such as O3, CO, and SO2. The study's findings underscore the appropriateness of support vector regression as a valuable instrument in projecting air quality metrics, highlighting its capacity to provide dependable insights into pollutant concentrations and contribute to a more nuanced comprehension of air pollution dynamics.

References

[1] Xiaoju Li-Luqman, C. A. Shafreeza, S, S, M. Siti, A. H. "Long-Term Air Pollution Characteristics and Multi-Scale Meteorological Factor Variability Analysis of Mega-Mountain Cities in the Chengdu-Chongqing Economic Circle", *Water Air Soil Pollut,* (2023): 228-334. https://doi.org/ 10.1007/s11270-023-06279-8

[2] Axel, G. M. M Younghak, M. Younghwan, Y and Jaehum, A. "Distributed Deep Features Extraction Model for Air Quality Forecasting", *Sustainability 12, no. 19* (2019): 01-19. https://doi.org/ 10.3390/su12198014

[3] Srivastava, C. Prakash Singh, A. Singh, S. "Estimation of Air Pollution in Delhi Using Machine Learning Techniques*: International Conference on Computing Power and Communication Techniques,* 7 (2018): 10-15. https://doi.org/ 10.1109/GUCON.2018.8675022

[4] Preeti Aggarwal and Suresh Jain. "Impact of Air Pollutants from Surface Transport Source on Human Health: A modelling and Epidemiological Approach" Environ Int, p. 146-57. 2015. https://doi.org/10.1016/j.envint2015.06.010

[5] Pooja Bhalgat, Sejal Pitale and Sachin Bhoite. "Air Quality Prediction using Machine Learning Algorithms." *International Journal of Computer Applications Technology and Research,* 8, no. 09 (2019): 367-370. https://doi.org/10.1016/j.ast.2012.02.005

[6] Vinoth Thanagaraju, Kothalam Krishnan Nagarajan. "A Detailed Analysis of Air Pollution Monitoring System and Prediction Using Machine Learning Methods." *International Journal on Recent and Innovation Trends in Computing and Communication* 11, no. 2 (2023): 51-58. https://doi.org/10.17762/ijritcc.v11i2s.6058

[7] Srinivasa, G. Yashiv, M. Khyati, H. Raahil, A. Valarmathi and Arulkumaran. "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis." *Journal of Environmental and Public Health,* (2013): 01-26. https://doi.org/10.1155/2023/4916267

[8] Suprateek Halsana. "Air Quality Prediction Model using Supervised Machine Learning Algorithms" *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,6, no.4* (2020): 190-201. https://doi.org/10.32628/CSEIT206435

[9] Lee, M. Lin, L, Chen, CY. "Forecasting Air Quality in Taiwan by Using Machine Learning". *Sci. Rep 10, no.4153* (2020):51-65. https://doi.org/10.1038/s41598-020-61151-7

[10] Fabiana Martine, C. Ales, P. Sara Silva and Leonardo, V. "A Machine Learning Approach to Predict Air Quality in California" *Complexity,6, no.4* (2020): 01-23. https://doi.org/10.1155/2020/8049504

[11] Doreswamy, Harishkuma, K.S. Yogesh, K. M. Ibrahim, G. "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models" *Procedia Computer Science 171, (*2020): 2057-2066. https://doi.org/10.1016/procs.2020.04.221

[12] Singh, J. K. and Goel, A. K. "Prediction of Air Pollution by using Machine Learning Algorithms" *7th International on Advanced Computing and Communication Systems, (*2021): 1345-1349. https://doi.org/10.1016/procs.2020.04.221

[13] Samaher Al-Janabi, Mustafa Mohammad, Ali-Sultan. "A New Method for Prediction of Air pollution based on Intelligent Computation" *Soft Computing 24, no.5, (*2020): 19-49. https://doi.org/10.1007/s00500-019-04495-1

[14] Aditya, C. R. Chandana, R. Nayana, D. K. Praveen Gandhi, V. "Detection and Prediction of Air Pollution using Machine Learning Models" *International Journal of Engineering Trends and Technology 59, no.4, (*2018): 204-207. https://doi.org/10.14445/22315381/IJEIT-V59P238

[15] Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou. "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization" *Big Data and Cognitive Computing 2, no.1, (*2018): 01-05. https://doi.org/10.3390/bdcc2010005

[16] Lan Gao, Changjie Cai, Xiao-Ming Hu. "Air Quality Prediction Using Machine Learning" *Wiley 11, no.1, (*2022): 10-15. https://doi.org/10.1002/9781119817512.ch11

[17] Sumitra Muniandy, Syuhaida, I, Md Ezamudin, S. "Revenue/Cost Production Sharing Contract (PSC) Fiscal Regime on Marginal Gas Fields in Malaysia: Case Study" *Progress in Energy and Environment 26, no.1, (*2023): 11-18. https://www.akademiabaru.com/submit/index.php/progee

[18] Gunasekar Thangarasu, Dominic, P.D.D. "Prediction of Hidden Knowledge from Clinical Database using Data Mining Techniques" *International Conference on Computer and Information Sciences, (*2014): 01-05. https://doi.org/10.1109/ICCOINS.2014.68684314

[19] Chin, Y. H. Terh, J. K. Zheng. Y. K. "Current Status of Green Building Development in Malaysia" *Progress in Energy and Environment 25, no.1 (*2023): 01-09. https://doi.org/10.37934/progee.25.1.19

[20] Yong Liu, Peiyu Wang, Yong Li, Lizia Wen and Xiaochao Deng. "Air Quality Prediction Models based on Meteorological Factors and Real-time data of Industrial Waste Gas" *Scientific Reports 12, no.9253 (*2022): 01-05. https://doi.org/10.1038/s41598-022-13579-2