Int. J. Nonlinear Anal. Appl. In Press, 1–5
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2025.34633.5179

Prediction of air quality index using machine learning algorithms: A case study of Tehran

Arshia Azizi, Sajjad Rahmany*

Department of Mathematics and Computer Science, Damghan University, Damghan, Iran

(Communicated by Madjid Eshaghi Gordji)

Abstract

Air quality index (AQI) forecasting is a useful tool for increasing the general public 's awareness of the state of the air in the next days. This is one of the most significant problems facing any country. In this study, machine learning algorithms are used to predict the AQI in Tehran. The six important regression models are applied to forecast AQI on a daily basis. Models were compared and evaluated using statistical measures such as Mean Absolute Error (MAE), coefficient of determination, and root mean square error (RMSE). Based on these evaluations, the best model was selected. ExtraTreesRegressor is thought to be the best model for forecasting AQI in all seasons based on its outcomes. The results demonstrate that the ExtraTreesRegressor 's determination coefficient is nearly 1, and that the values of MAE and RMSE are respectively 0.002 and 0.004.

Keywords: air pollution, air quality index, machine learning, algorithms, mean absolute error, root mean square error
2020 MSC: 62P12, 68T05, 86A10, 92C40

1 Introduction

Tehran is one of several cities throughout the world that are affected by air pollution, which is a serious environmental issue. Recent years have seen a considerable decline in air quality due to the city's rapid population increase, heavy traffic, and industrial activity. In order to assist authorities in making sound decisions and executing the necessary steps to enhance the quality of the air, there is an urgent need for efficient monitoring and prediction systems [3, 6]. Recently, many researchers have focused on forecasting air pollution using machine learning [10], neural networks [3, 12], and deep learning [6]. Machine learning techniques are widely used in environmental sciences, including weather forecasting, soil erosion, waste management, dust storms, and air pollution [1]. Conventional air pollution prediction techniques can be divided into statistical methods , artificial intelligence, and numerical forecasting [2]. Sharma et al. [11] time-series analysis of data from 2009 to 2017 was used to predict the air quality in New Delhi. To create a forecasting model based on deep learning, Kaya and Oguducu [7] used PM10 hourly data from Istanbul (Turkey) between 2014 and 2018. Gocheva-llieva et al. [5] developed a model for daily prediction that had 90% accuracy using the classification and regression tree technique.

The rest of the paper follows the materials and methods in Section 2, the results including the data preparation and refinement and air pollution prediction are presented in Section 3, and conclusions is presented in Section 4.

*s_rahmani@du.ac.ir
Email addresses: arshiyaazizi1999@gmail.com (Arshia Azizi), s_rahmani@du.ac.ir (Sajjad Rahmany)

Received: July 2024 *Accepted:* January 2025

2 Materials and Methods

The data used in this paper consist of daily air quality data from Tehran Air Quality Control Company (AQCC) from several monitoring stations across Tehran from March 21, 2020, and March 21, 2023. After researching and reading several articles about Tehran's air pollution, the following 14 features have been selected in Table 2 [9, 8, 14, 13, 4].

Symbol	Feature
PM	Particulate Matter
NO2	Nitrogen Oxides
SO2	Sulfur Dioxide
CO	Carbon Monoxide
O3	Ozone
Temperature	A physical quantity known as temperature expresses quantitatively how hot or cold something is
Humidity	The concentration of water vapor in the air is known as humidity
Precipitation	Precipitation is any byproduct of atmospheric water vapor condensation that falls from clouds as a result of gravitational pull
Wind Gust	A wind gust is a momentary increase in wind speed
Wind Speed	The most important component of the atmosphere is wind speed, which is the rate at which air shifts from high to low pressure
Sea Level Pressure	Pressure within the atmosphere of Earth
Visibility	The measurement of the distance at which a light or item can be seen clearly
Solar Radiation	Solar irradiance is the surface power density of electromagnetic radiation that is received from the Sun in the wavelength range of the measuring device
UV Index	The ultraviolet index, or UV index, is a globally recognized indicator of the amount of UV light that can cause sunburns at a specific location and time

All of the data used in this work was normalized as scaled to range (0,1) in order to ensure that all numerical values were on the same scale and that large values did not dominate smaller ones. Figure 1 shows the general structure of the suggested model.

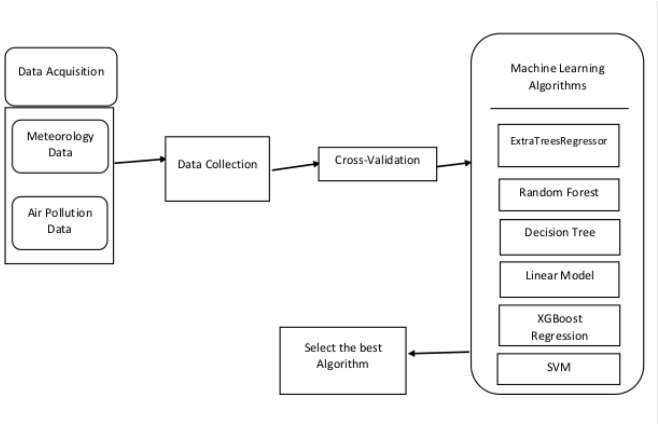

Figure 1: The proposed best air pollution prediction model.

The performance of our approach was confirmed using the 10-fold cross-validation method. Applying machine learning algorithms like ExtraTreesRegressor, Random Forest, Decision Tree, Linear Model and XGBoost Regression will be done in the coming steps. In order to evaluate the prediction performance of the proposed model, we used the following measure:

• The mean absolute error is the average absolute difference between the predicted value and actual value, and is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - x_i|}{N}.$$

• The root mean square error is the square root of the distance between the predicted and actual value:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\text{Predicted}_i - \text{Actual}_i)^2}{N}}.$$

• The coefficient of determination, sometimes called coefficient, is the fraction of the variation in the dependent variable that is predicted from the independent variable(s), denoted R^2:

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Our results show that the ExtraTreesRegressor model performed well in predicting AQI. The most important variables for predicting AQI concentration were found to be temperature, humidity, wind speed, and traffic volume.

The ExtraTreesRegressor algorithm is a variant of the popular Random Forest method, which uses an ensemble of decision trees to make predictions. The ExtraTreesRegressor algorithm adds an additional level of randomness to the decision tree construction process, resulting in improved performance and faster training times.
The most important variables for predicting AQI concentration were found to be PM2.5, PM10, O3, Visibility, NO2 and traffic volume. Figure 2 shows the relative importance of each variable in the model.

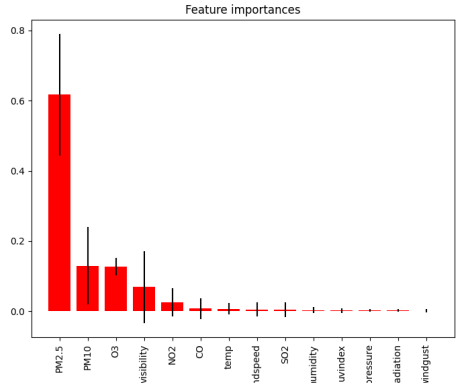

Figure 2: Relative importance of each variable in predicting AQI concentration using the ExtraTreesRegressor algorithm.

3 Results

We evaluated the performance of the model using several metrics, including mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and coefficient of determination (R^2). The mean absolute error (MAE) measures the difference in errors between paired observations describing the same occurrence. The results of the algorithms applied to the data are shown in Table 3 and Figure 3 shows a comparison of the predicted and actual AQI concentrations for the testing set.

Algorithm	MAE	MSE	RMSE	R^2
ExtraTreesRegressor	0.002	1.94	0.004	0.996
Random Forest	0.002	0.0001	0.010	0.97
Decision Tree	0.002	4.84	0.006	0.99
Linear Model	0.015	0.0005	0.02	0.91
XGBoost Regression	0.003	3.43	0.005	0.99
SVM	0.04	0.002	0.046	0.62

Overall, our results suggest that the ExtraTreesRegressor algorithm can be a useful tool for predicting air pollution in Tehran. The default setting for the ExtraTreesRegressor algorithm used in this study is 100 trees. The ExtraTreesRegressor algorithm produces the results shown below by altering the number of trees:

Number of Trees	MAE	MSE	RMSE	R^2
100	0.002	2.23	0.004	0.993
500	0.002	2.35	0.004	0.995
1000	0.002	1.94	0.004	0.996

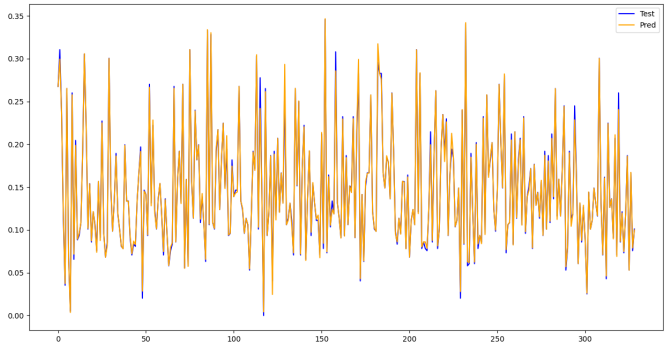

Figure 3: Comparison of predicted and actual AQI concentrations for the testing set using the ExtraTreesRegressor algorithm.

4 Conclusions

The focus of the research conducted in this article was on predicting air quality in Tehran by utilizing machine learning algorithms. Specifically, the ExtraTreesRegressor algorithm was employed to predict the concentrations of air pollutants in the city, based on various environmental and meteorological factors. Our study's findings indicated that the ExtraTreesRegressor algorithm was successful in predicting air pollutant concentrations, with an overall accuracy rate of 99% for predicting pollutant AQI.

One of the major advantages of our study was the use of a comprehensive and novel dataset that encompassed various meteorological and environmental factors. This allowed us to identify the crucial factors responsible for air pollutant concentrations and develop an accurate model to predict pollutant concentrations. Our research also highlights the potential of machine learning algorithms in predicting air pollutant concentrations, which can be leveraged to inform public health policies and decrease the adverse effects of air pollution on public health.

However, it's worth noting that our study's scope was limited to air pollutant concentrations in Tehran, which may not be generalizable to other regions or cities. Additionally, our research did not consider the impact of human behavior and activity patterns on air pollutant concentrations, which could be an important factor to consider in future studies.

In conclusion, our study provides crucial insights into the potential of machine learning algorithms for predicting air pollutant concentrations, emphasizing the necessity for further research in this domain. By enhancing and refining these algorithms, we can gain a better understanding of the factors contributing to air pollution and develop more effective approaches to mitigate its negative impacts on public health.

References

[1] S. Agarwal, S. Sharma, M.H. Rahman, S. Vranckx, B. Maiheu, L. Blyth, S. Janssen, P. Gargava, V.K. Shukla, and S. Batra, *Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions*, Sci. Total Envir. **735** (2020), 139454.

[2] L. Bai, J. Wang, X. Ma, and H. Lu, *Air pollution forecasts: An overview*, Int. J. Environ. Res. Public Health **15** (2018), no. 4, 780.

[3] M.R. Delavar, A. Gholami, G.R. Shiran, Y. Rashidi, G.R. Nakhaeizadeh, K. Fedra, and S. Hatefi Afshar, *A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran*, ISPRS Int. J. Geo.-Inf. **8** (2019), no. 2, 99.

[4] Z. Ghaemi, A. Alimohammadi, and M. Farnaghi, *LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran*, Environ Monit Assess **190** (2018), no. 5, 300.

[5] S.G. Gocheva-Ilieva, D.S. Voynikova, M.P. Stoimenova, A.V. Ivanov, and I.P. Iliev, *Regression trees modeling of time series for air pollution analysis and forecasting*, Neural Comput. Appl. **31** (2019), no. 12 9023–9039.

[6] A. Heydari, M. Majidi Nezhad, D. Astiaso Garcia, F. Keynia, and L. De Santoli, *Air pollution forecasting*

application based on deep learning model and optimization algorithm, Clean Technol. Envir.Policy **24** (2022), no. 2, 607–621.

[7] K. Kaya and Ş. Gündüz Öğüdücü, *Deep flexible sequential (DFS) model for air pollution forecasting*, Sci. Rep. **10** (2020), no. 1, 3346.

[8] H. Luo, Q. Guan, J. Lin, Q. Wang, L. Yang, Z. Tan, and N. Wang, *Air pollution characteristics and human health risks in key cities of northwest China*, Sci. Total Envir. **269** (2020), 110791.

[9] R. Munsif, M. Zubair, A. Aziz, and M.N. Zafar, *Industrial air emission pollution: Potential sources and sustainable mitigation*, Environ Emissions, Chapter 4, 2021.

[10] A.K. Rad, R.R. Shamshiri, A. Naghipour, S.O. Razmi, M. Shariati, F. Golkar, and S.K. Balasundram, *Machine learning for determining interactions between air pollutants and environmental parameters in three cities of Iran*, Sustainability **14** (2022), no. 13, 8027.

[11] N. Sharma, S. Taneja, V. Sagar, and A. Bhatt, *Forecasting air pollution load in Delhi using data analysis tools*, Procedia Comput. Sci. **132** (2018), 1077–1085.

[12] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu, and T. Chi, *A novel spatiotemporal convolutional long short-term neural network for air pollution prediction*, Sci. Total Envir. **654** (2019), 1091–1099.

[13] L. Wu, N. Li, and Y. Yang, *Prediction of air quality indicators for the Beijing-Tianjin-Hebei region*, Cleaner Prod. **196** (2018), 682–687.

[14] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, *PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing datat*, Atmosphere **10** (2019), no. 7, 373.