International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR AIR QUALITY FORECASTING IN MAHARASHTRA

Miss. Jayashree Bhuskute[1*], Dr. Ashok Tayade[2]

[1*]Department of Statistics Dr. BAMU, Chh.Sambhajinagar 431004(M.S) India
statistics.jsb@bamu.ac.in
[2]Department of Statistics Dr. BAMU, Chh.Sambhajinagar 431004(M.S India aytayade@gmail.com

Abstract

This paper addresses the machine learning (ML) based prediction of air pollution in various cities of the state Maharashtra. To conduct the study data collected by the Maharashtra Air Quality Monitoring Network, were used to overcome data scarcity and train the models. A comparison of PM2.5, PM10, nitrogen dioxide (NO2), sulfur dioxide (SO2), and ozone (O3) over the whole year of 2023 provides a clear picture of how the air quality index (AQI) changes over time. The most accurate models were the extreme gradient boosting (XGBoost) model (98.83%) and the naive Bayes classifier (NBC) (95.58%). The support vector machine (SVM) reported the lowest accuracy of 81%, which is still improving. This research study, in addition to illustrating how ML-based models can help predict future air quality trends, also highlights the advantages of using synthetic data to increase the accuracy of the predicted outputs. These results provide a useful perspective for policymakers in the formulation of effective interventions that can be implemented to alleviate the air pollution problem and ensure that the urban centers in Maharashtra are transformed into places where people can live quality lives. This article establishes the relevance of further research on the use of synthetic data and machine learning in the fight against poor air quality.

Keywords: Air quality index (AQI), Machine learning (ML), XGBoost, Naive Bayes classifier (NBC), Particulate matter (PM).

Introduction

The primary obstacle impeding development in developing nations, such as India, is contamination of the atmosphere. An increase in the quantity of air pollution (AP) accompanied by an increase in the density of established land use is a matter of great concern (Q. Weng & S. Yang., 2006). Therefore, in order to ensure that the "Air Quality (AQ)" remains within the acceptable thresholds according to established criteria, it is necessary to collect timely and regular data on fluctuations in the level of air pollution in metropolitan areas must be collected (Vashisht, Somvanshi, & Shrivastava, 2018) Approximately 0.62 million premature excess deaths cases in India are attributed to outdoor air pollution, making it the fifth most common cause of death in 2012 (Rathi & Rathod, 2016) (NYT, 2014) , followed by hypertension, polluted indoor air, smoking cigarettes, and inadequate nutrition. (Patil, Radhika, & Dinde, 2023) In their study of different regions of India reported that industries contribute to even higher levels of air pollution in different regions of India than residential areas.

Modern society continues to address the grapple with a grave issue of air pollution. According to the consensus of experts, socioeconomic human activities exert the greatest influence on the environment (Priya & Sathya, 2019) (Zhang, Chen, & Wu, 2019). Throughout most of the year, numerous industrialized and developing towns suffer from substandard air quality. Ground travel, industrial emissions, and anthropogenic behaviors are the primary contributors to air pollution (OnkalEngin, Demir, & Hiz, 2004). Exposure to diffuse APs in urban centers across industrialized and developing countries significantly affects the quality of life of the population. A previous study (Schwela & Dietrich, 2000) indicated that almost 1.5 billion individuals are exposed to elevated levels of "suspended particulate matter (SPM)", "sulphur dioxide (S02)", and ozone in the ambient air. As

Received: August 06 , 2025

Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

described in the study (Kumari, Somvanshi, & Zubair, 2021) have employed research methodologies to analyze the spatial and temporal characteristics of PM_{10} and $PM_{2.5}$ and the trace gases O3, “nitrogen dioxide (NO2)” and “carbon monoxide (CO)” in Mumbai with evidence of a diabolical rise in pollution levels (TAYADE, 2012).

This study reveals the importance of intervention measures focused on industrial areas in order to control pollution successfully. The health consequences resulting from pollutants in the air incurred a financial burden of almost USD 81 billion in 2011, which is comparable to 5.7% of India's "Gross Domestic Product (GDP)" (Mangalekar, Jadhav, & Raut, 2015) (Shama, Taneja, & Bhatt, 2020)compared with the pollution level among the states ,which shows that, owing urbanization, the pollution level increases and directly impacts the quality of human life in the states such as Karnataka, Maharashtra, West Bengal and the Delhi NCR. This leads to a considerable burden on the cardiovascular system, respiratory system, and movement of adults (WB, 2013;2015) (Chauhana, et al., 2020). While several earlier studies have examined AQ in various metropolitan areas in India by (Agarwal, Aviral, Kaushik, Kumar, & Mishra., 2020), none of these studies have explicitly analyzed and released liquid and particle amounts for every phase of the study to provide a comprehensive assessment of the AP situation in Sangali city.

Economic development was found to be positively associated with the growth of air quality monitoring networks by (Sharma & Kota, 2024)**.** Thus, they reported a reduction in SO2 concentrations and relatively stable NO2 levels but with occasional increases. The detrimental impacts of ambient air pollution on human health are well recognized in industrialized countries (Purohit, Chauhan, Vyas, Vyas, & Sing, 2017). Some of the current studies (Sonavane & Pinjar, 2024) have analyzed the levels of ambient air quality in Sangali city and highlighted the need to employ appropriate monitoring tools for the prevention and control of pollution in the city. Unfortunately, the general population is inadequately educated about the detrimental impacts of AP on human health, particularly in developing countries where coal is extensively used for combustion and where the automobile population is increasing (Kulkarni, Muley, Deshmukh, & Bhalchandra, 2018). Only more recently has environmental AP started be recognized as a chronic concern for all residents of urban regions in developing countries. Similarly, (Das & Ghosh, 2023) confined the study to Maharashtra’s nonattainment cities and advised measures for NCAP to achieve a pollution cutoff of 20–30% by 2024.

In general, the air quality index (AQI) is a useful instrument for elucidating the conditions of the surrounding air. The process condenses the intricate data regarding various air pollutants into a singular numerical value referred to as the index value, accompanied by identification and color (N. K. Rai & Vyas, 2017). Another work by (Khedekar & Thakare, 2023) also developed a connection between airborne pollutants and meteorological factors via the use of fine-grained data for the Pune region. The two reports they used statistical methods to analyze massive datasets constantly collected for environmental purposes to understand the trends of air pollution (Sonar & TAYADE, 2015).

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

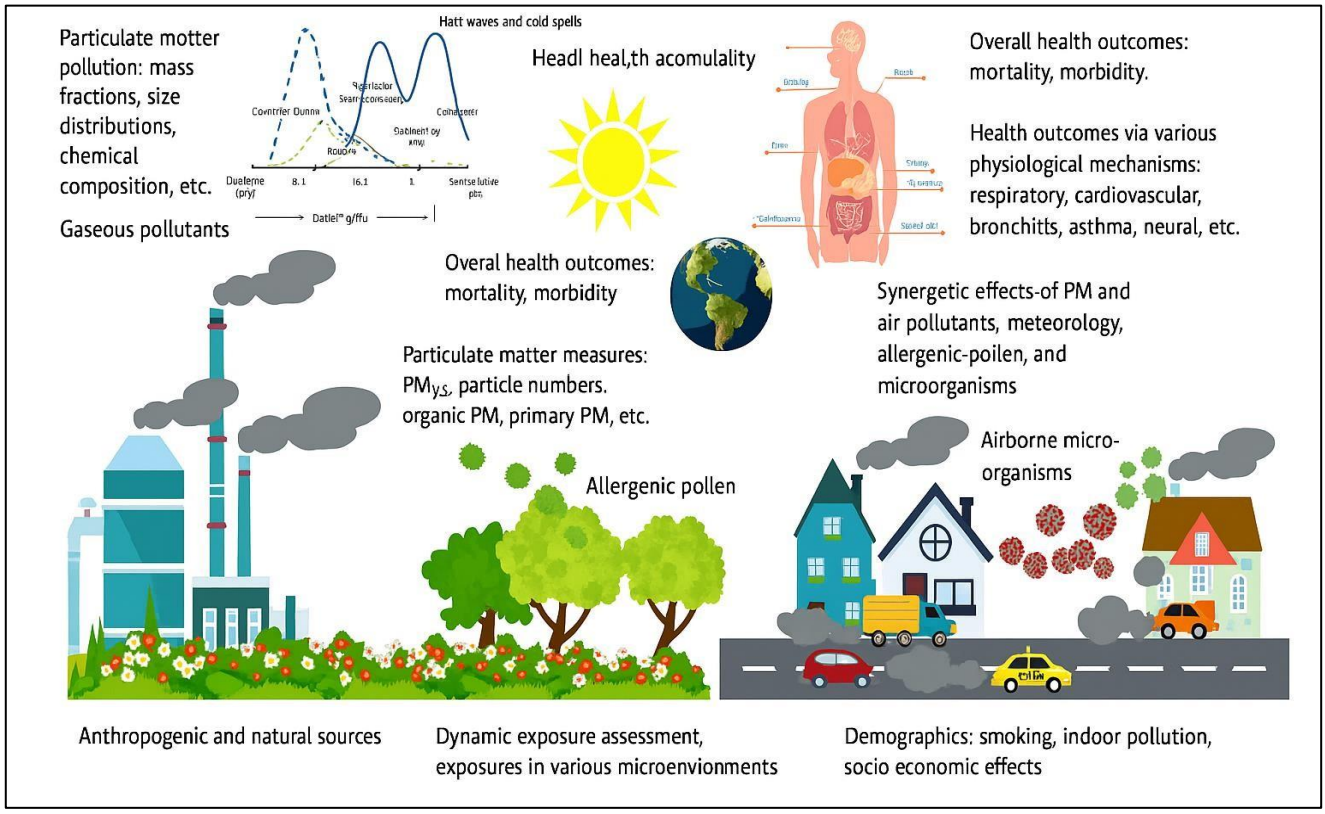


Figure 1 A schematic diagram that illustrates the primary aspects involved in evaluating the exposure to air pollution

Metropolises are seeing significant environmental issues in AP, large humanity, demand for electricity, numerous industries, and extensive car fleets. Therefore, urban areas are often considered high-risk zones where the human population is susceptible to adverse health effects caused by air pollution, such as "cardiovascular disease (CD)", respiratory disease (RD), and "chronic obstructive pulmonary disease (COPD)" (Butler, et al., 2008) (S, 2006) (Molina & Molina, 2004). In the study of (Pathan, 2022) have discussed various health outcomes associated with air pollution and stated that future research should focus on the different health effects of air pollution which is fatal and can lead to different severe diseases such as cardiovascular diseases, diabetes, obesity, and cancer. A thorough calculation of this risk is necessary to assist air quality authorities in improving the sustainability of urban living.

Machine learning (ML)-based AQI forecasting algorithms were demonstrated to be more dependable and consistent. Sophisticated technology and sensors facilitated the streamlined and accurate gathering of data. A comprehensive analysis of the significance of supervised machine learning algorithms for applied environmental protection problems was conducted by (Al-Jamimi & Saleh, 2019). This study examines AP data from Indian cities and evaluates twelve air contaminants and their corresponding AQIs. Initial preprocessing and cleaning of the dataset is followed by the application of data visualization techniques to enhance insights and explore concealed emerging trends and patterns. This work leverages the fundamental concept of coefficients of correlation in ML models, a topic that has received limited attention from writers in the existing literature (Alade, Rahman, & Saleh, 2019). Widely used ML models are evaluated within the framework of this reproducing approach. The achievements of these methods are thereafter evaluated via conventional quantitative measures. Numerous experts in the field and certain authors of ML applications, such as (Ayturan, et al., 2020), employ these metrics. “Particulate matter” (PM), which is categorized on the

basis of its lightweight diameter, is a hazardous contaminant that has detrimental impacts on human health.

While there have been developments in predicting the different patterns of air pollution levels via ML approaches, there is a relative research gap in case-by-case application of the models in the functional regions of Maharashtra. Previous research very often covers regional analysis only or does not consider some essential factors, such as land use, socioeconomic conditions and actual weather conditions that may have an enormous impact on the concentration of pollutants (Kulkarni & Jadhav, 2025) in the air. Additionally, studies that compare the performance of distinct categories of machine learning algorithms under different circumstances pertinent to cities in Maharashtra are relatively rare. Addressing these gaps could improve the precision and relevance of forecasts of air pollution, which may in turn improve policy and health interventions.

This research stands out within the current work on ambient air pollution in developing regions, including Maharashtra, by focusing on its effects on public health and the urban environment. This study provides a systematic view of how air pollution varies across the state, offers recommendations for effective air quality management, and supports policy development aimed at reducing environmental harm. This study examines how several machine learning models can predict the air pollution levels different cities and determine the best model that can be used in the future. It also examines improving the feature selection approach based on advanced optimization methods, Harris hawk optimization (HHO), and compares the predictive performance of models, extreme gradient boosting (XGBoost), the naive Bayes classifier (NBC) and the support vector machine (SVM), on observed air quality data. This study further searches for spatial and temporal trends in air pollution incidence and the effects of environmental variables. land surface temperature (LST) in terms of predictive efficacy, especially in terms of precision and recall. Through the combination of these factors, this study will contribute to methodological developments in air quality prediction and generate useful knowledge that can guide interventions in environmental health in the vast array of urban environments in Maharashtra.

Methods and Materials

The steps involved in developing the methodology section. First, data concerning the geographical and topographical features of Maharashtra are collected to understand the environment of the region. The AQ data used in the study were obtained from the Maharashtra Air Quality Monitoring Network (MAQMN) for particulate matter such as PM2.5, NO2, and O3. Preprocessing involves characteristic selection via Harris hawk optimization (HHO), where after data cleaning is conducted via mean and median imputation, normalization is performed via MinMaxScaler. The air quality index (AQI) is derived mathematically from the concentration of pollutants and the standard allowed for them. “Naive Bayes Classifiers (NBCs)” and “support vector machines (SVMs)” are then applied to the data to do the same. The models are trained and evaluated, and their effectiveness is compared on the basis of parameters such as accuracy, precision, recall and the F1 score

Study Area

Maharashtra is a state in western India that is one of the largest states in the country, with a population of more than 124 million recorded in the 2021 census, and occupies an area of approximately 307,713 sq. km (118,809 sq. miles), which is approximately 8.9% of the total population in India. Geographically, the state is endowed with many physical features, including the extensive plains of basaltic lava shaped by events generated previously, the Western Ghats range, which forms the western part of the state, the Vindhya and Saputra ranges, and the northern region, which contains high-value minerals such as sandstone and bauxite. The air quality index (AQI) monitoring stations were established by the Maharashtra Pollution Control Board (MPCB) under the National Air Monitoring Program (NAMP) in all major cities and towns, including Mumbai, Pune, Nagpur, Nashik,

Aurangabad (Chhatrapati Sambhajinagar), Solapur, Amravati, Akola, Kolhapur, Jalgaon, and Chandrapur, by associated autonomous institutes directly connected with the MPCB. Geographical and topographical data related to the study region were also acquired via official data (maharashtra.gov.in) so that the various regions of India that have diverse landscapes as well as populations could have an ideal representation of the geographical profile of the state in the context of research (www.census2011.co.in, 2011).

Dataset Description

In analyzing the trend of the ambient air pollution for several cities in Maharashtra, the "Maharashtra Air Quality Monitoring Network (MAQMN)" (www.mpcb.gov.in, n.d.) (www.aqi.in, n.d.) is a pragmatic source of datasets. This dataset contains all-round data regarding the various indicators of air quality obtained from different monitoring stations of the state. MAQMN contains data on a range of important air pollutants, including particulate matter-Pm2.5 and PM10, NO_2, SO2, and O3, which are collected at different time scales. The dataset provides the most valuable basis for detecting air pollution patterns, determining zones of elevated risk, and for evaluating the management and regulation of air quality. This dataset enables the study of temporal and spatial variations in pollution levels, thus providing useful insights into the environmental health effects of cities in Maharashtra.

In our study, we first focused on selecting the most important features to improve model performance. For this purpose, we used the Harris hawk optimization (HHO) technique, which works like the hunting strategy of hawks searching widely at first and then zooming in on the best targets. To handle missing data, we replaced gaps with either the mean or median of the available values, with the median being more reliable when extreme values are present. Finally, we applied MinMax scaling to bring all the features into the same range (0-1), which helps the model treat every feature fairly without distorting the original data patterns.

Methodology Field Data Collection

The air pollution index (API) in this study was calculated via ground-level concentrations of $PM_{2.5}$, PM_{10}, NO_2, CO, and O_3, which were obtained from the SAFAR network established by the Indian Department of Science and Technology. Ten sites in Mumbai were chosen for monitoring, The API formula uses the pollutant concentrations estimated to their limits in CPCB standards and the computations of individual pollutant indices by (Rao, Hima, G, Indracanti, & Anjaneyulu, 2004)

$$API_{Pollutant} = \left(\frac{Pollutant}{S_{pollutant}}\right) \times 100 \quad (1)$$

The overall API was then derived by averaging the normalized values of the selected pollutants. An AQI scale was used for classification, ranging from optimal atmospheric conditions (0–25) to intense contamination (>100) was used for classification.

Table 1 AQI scale for ratings

Sr. No.	AQI rate	Observations
1	00-25	Optimal atmospheric
2	36-50	AP of low intensity
3	51-75	Moderate levels of AP
4	75-100	High levels of AP
5	>100	Intense atmospheric contamination

Land surface temperature (LST)

The LST was derived from Landsat 8 satellite data via the methodology outlined in the Space Data Users Handbook. The thermal band digital number (DN) values were converted to spectral radiance (Lλ) values and subsequently to brightness temperature (TB) values via calibration constants. Emissivity correction, which is based on land cover type, was applied following established

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

approaches (Artis and Carnahan,1982; Weng et al,2004). The corrected LST was then computed considering the wavelength-dependent atmospheric effect (Snyder, Wan, Zhang, & Feng, 1998).

Model selection

Machine learning algorithms were evaluated for air quality prediction:

- XGBoost – A gradient boosting method optimized for high efficiency and accuracy.
- Naive Bayes Classifier (NBC) – A probabilistic model based on Bayes' theorem with independence assumptions among features.
- Support vector machine (SVM) – A supervised learning model in which users kernel functions to optimize decision boundaries for classification or regression.
- Artificial Neural Network (ANN): A learning model built on the human brain with a network of neurons, which is used to learn complicated patterns on the basis of the data to make accurate predictions.
- K-nearest neighbors (KNN): A straightforward instance-based learning algorithm that which makes predictions about whether a given case belongs to the majority class or the average of the nearest training instances.

Training, Testing and Evaluation

The collected data were split into training and testing sets. Model performance was assessed via metrics derived from the confusion matrix. Accuracy, precision, recall and f1-score. This matrix evaluated the ability of the model to generalize to unseen data and provided insights its predictive reliability.

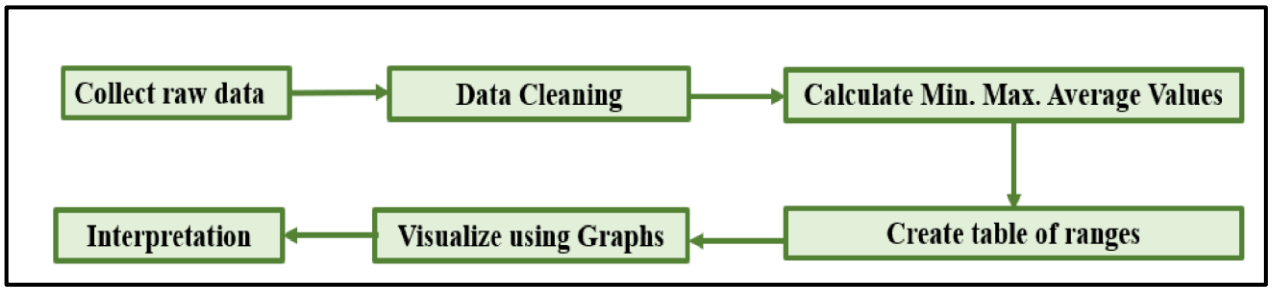


Figure 2 Flow chart of data cleaning

Validation

To validate the model developed for analyzing the AAQ in Maharashtra, employs an integrated methodology is employed. To check the reliability of the model due to changes in the input variables, sensitivity analysis is also performed to determine the effects of changes in PM2.5, PM10, NO2, SO2, and O3 on the AQI. In addition, model validation is performed via the performance of the NBC and the SVMs, where a set of test data, which is not used for training the models, is used to evaluate the accuracy of the models in making their predictions. The validation process also includes the computation of performance metrics such as accuracy, precision, recall and the F1score, which reflects how good or bad a given model is. This overall validation strategy corroborates and supports the reliability and utility of the methodology participate in estimating the concern and fluctuation of air excellence in individual cities in Maharashtra.

.

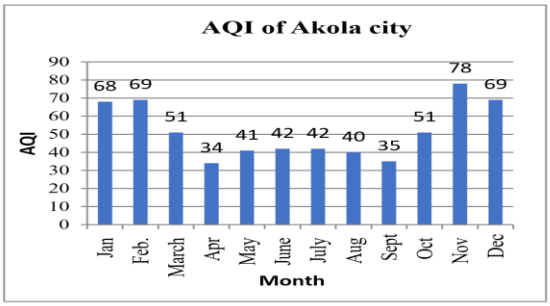
Received: August 06 , 2025

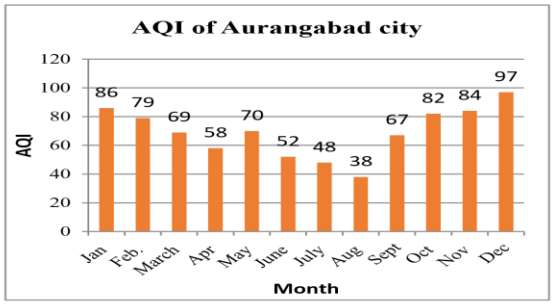
International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

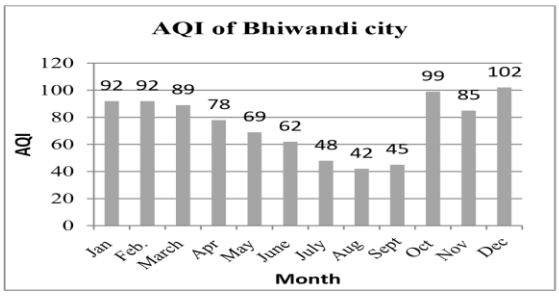
Table 2 Avg AQI of several cities in Maharashtra for the year 2023.

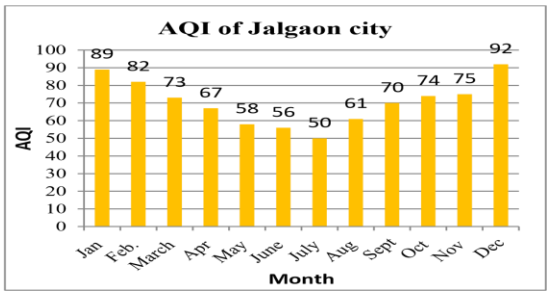
Avg Air Quality Index (AQI) 2023												
City	**Jan**	**Feb**	**March**	**Apr**	**May**	**June**	**July**	**Aug**	**Sept**	**Oct**	**Nov**	**Dec**
Akola	68	69	51	34	41	42	42	40	35	51	78	69
Aurangabad	86	79	69	58	70	52	48	38	67	82	84	97
Bhiwandi	92	92	89	78	69	62	48	42	45	99	85	102
Jalgaon	89	82	73	67	58	56	50	61	70	74	75	92
Mumbai	95	86	76	62	56	72	67	35	41	81	106	98
Nagpur	101	98	90	75	73	69	28	35	36	89	91	92
Nashik	89	78	70	68	64	71	18	24	36	73	65	87
Pune	98	102	96	65	60	32	28	30	39	82	63	78
Thane	95	98	76	80	72	78	26	20	71	94	81	93
Solapur	65	58	48	42	56	75	25	20	18	17	78	80

Table 2 shows the Avg AQI for different cities in Maharashtra for the year 2023, with monthly variations in the air pollution level. Therefore, cities such as Bhiwandi and Mumbai reported higher AQIs escalating in winter, indicating poor air quality, whereas Solapur presented comparatively lower AQI levels throughout the year. For example, December and January had greater pollution in all the cities, possibly because of temperature inversions. The collected AQI data can be mined via machine learning approaches to estimate future trends in air quality, which can serve as pointers for cities and policy makers to make relevant adjustments that can help improve and enhance residents' health. On the basis of variations and fluctuations in climatological conditions and pollution discrepancies between cities, the machine learning model can identify potential pollution surges and prompt the timely application of remedial measures.









Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

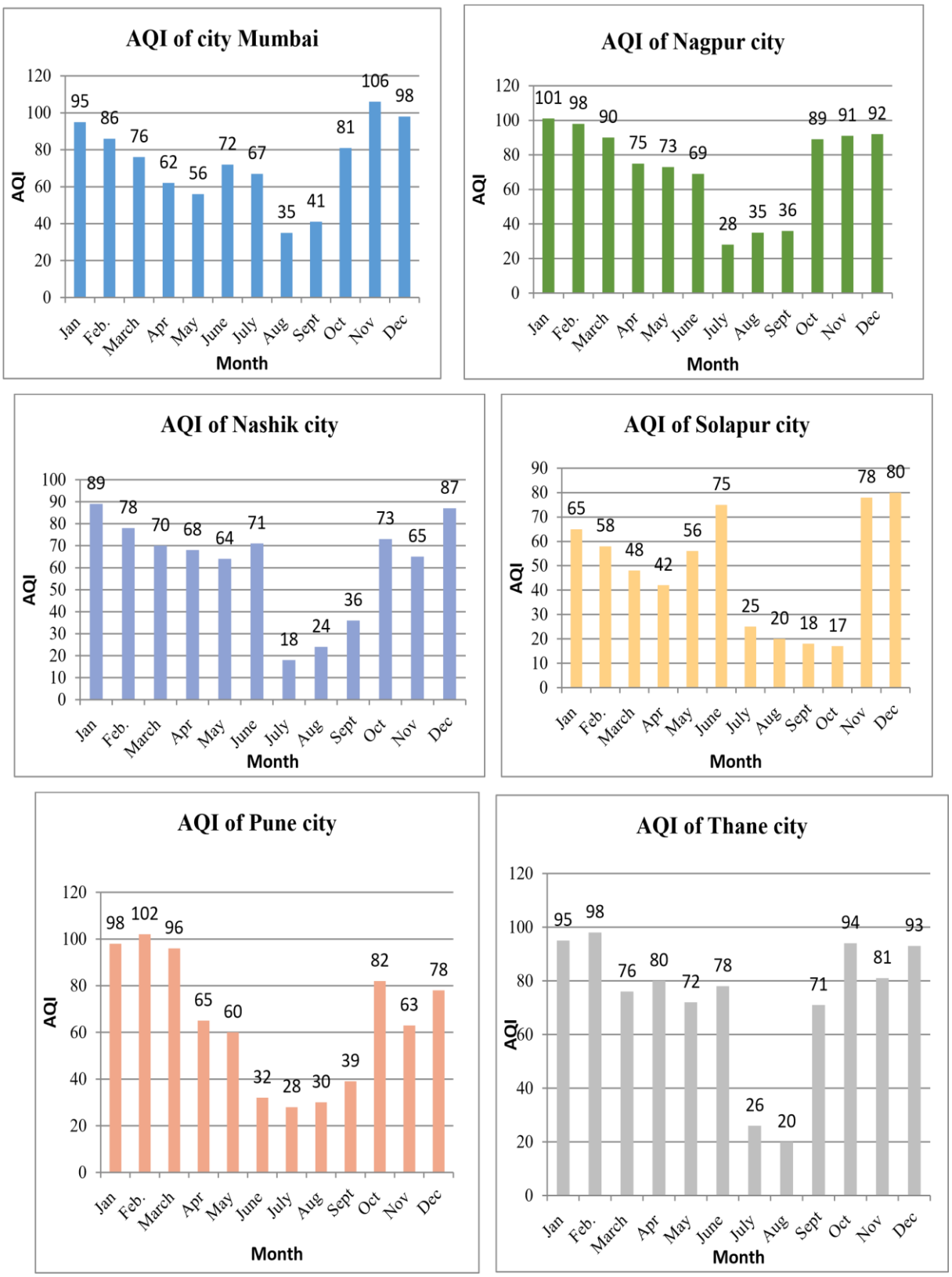


Figure 3 Fluctuations in AQI levels among various cities in Maharashtra

Figure 3 depicts the changes in the AQI for the cities of Maharashtra for the entire year, namely, Akola, Aurangabad, Bhiwandi, Jalgaon, Mumbai, Nagpur, Nashik, Solapur, Pune, and Thane. Moreover, each city clearly has a different trend in the AQI in different months of the year, indicating

Received: August 06 , 2025

that there are regular fluctuations in air quality. For example, Bhiwandi and Solapur have a comparatively higher AQI in the last quarter of the year only, and Mumbai has slightly increased throughout the year. However, the AQI in Nagpur and Aurangabad markedly increased toward the later months of the year. These variations offer valuable information that can be used to generate data for further machine learning for the prediction of degrees of pollution. With reference to these predictions, it is possible to examine pollution trends in various areas and use them for the development of effective prevention measures concerning air pollution.

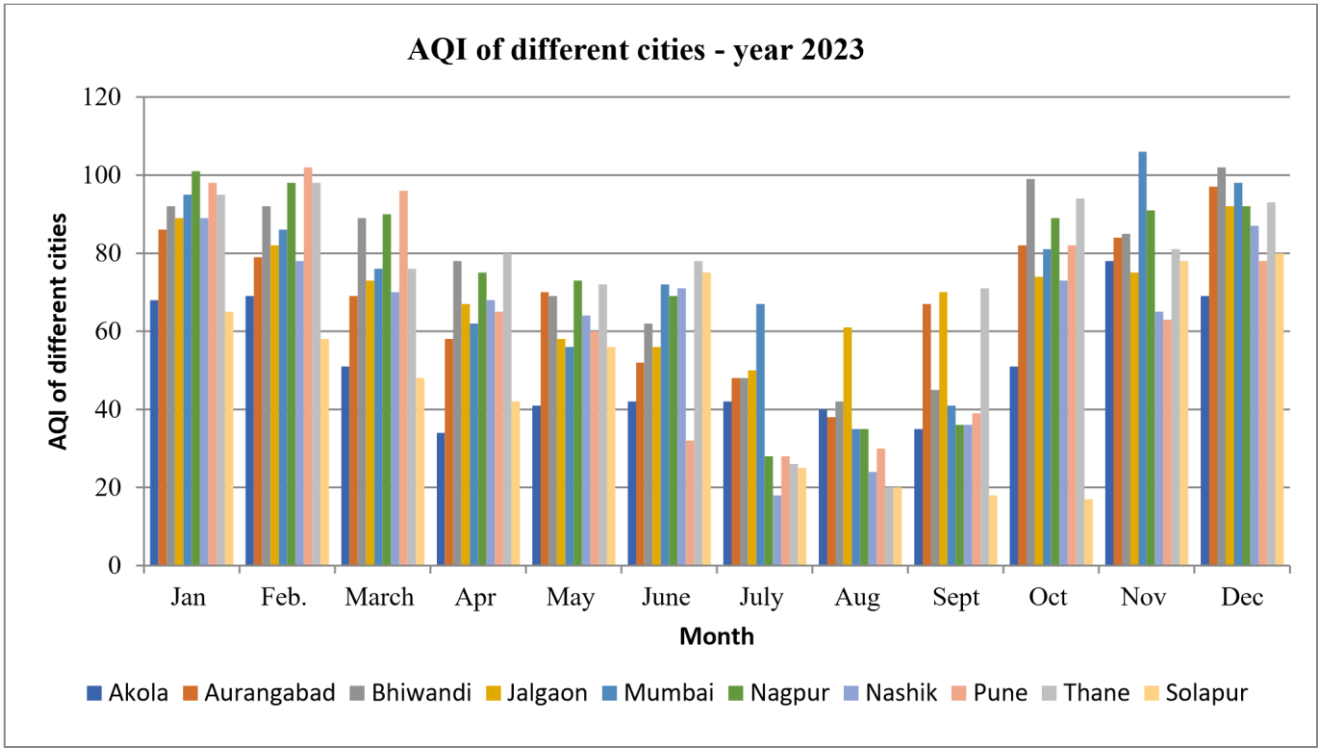


Figure 4 Depiction of the AQI of different cities in Maharashtra

Figure 4 shows a comparison of the AQI for the cities mentioned in Table 2. The temporal variation in the AQI for the year 2023 is shown, which shows that among all the cities, Mumbai has the highest AQI index in November, and Solapur has the minimum AQI index in both September and October. Significantly, the lowest AQI index is noted or expected from July -September; hence, the climatic conditions of those months are deemed beneficial to the air quality. These forecasts will enable the determination of pollution trends in various districts and the development of workable measures to stop air pollution.

Table 3 Monthly average AQI and key pollutant levels in Maharashtra for 2023.

Month	AQI (Average)	PM2.5 (µg/m³)	PM10 (µg/m³)	SO2 (µg/m³)	Ozone (µg/m³)	Temperature (◦C)
Jan	60 (Moderate)	80	40	5	18	20
Feb	55 (Moderate)	75	35	6	16	22
March	70 (Moderate)	90	45	5	14	25
Apr	85 (Unhealthy)	110	55	7	23	30
May	90 (Unhealthy)	120	60	8	25	33
June	50 (Good)	60	30	4	13	28

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

July	45 (Good)	55	25	3	12	26
Aug	55 (Good)	65	28	4	11	27
Sep	67 (Moderate)	22	52	2	14	42
Oct	75 (Moderate)	95	45	6	16	30
Nov	80 (Unhealthy)	110	50	7	24	25
Dec	65 (Moderate)	80	40	5	17	22

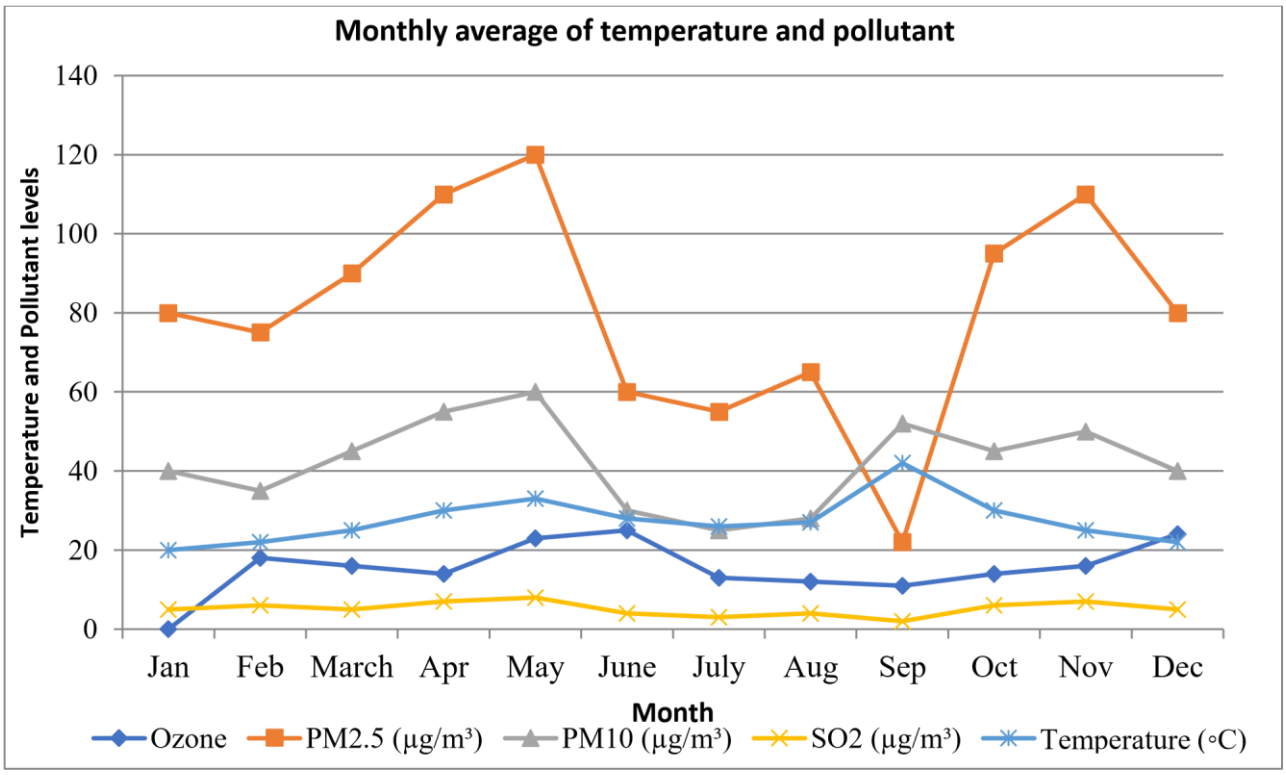


Figure 5 Monthly patterns of air contaminants (PM2.5, PM10, ozone and SO2) and temperature in Maharashtra

In Table 4 and Figure 5, various air quality parameters are displayed on a monthly basis, along with the concentrations of ozone, PM2.5, PM10 and SO2 along with temperature changes throughout the year. On average, ozone levels range between 0 and 25 µg/m³, with the highest being 22 µg/m³ in June. PM2.5 is relatively high, varying from 20 to 120 µg/m³ and peaking at approximately 120 µg/m³ in May. PM10 fluctuates slightly at approximately 20–60 µg/m³. SO2 remains low at an average of 10 µg/m³ until the end of the study. The temperature is generally seasonal, ranging from approximately 20°C in the winter months of January and February to approximately °C40 in August before it begins to fall again.

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

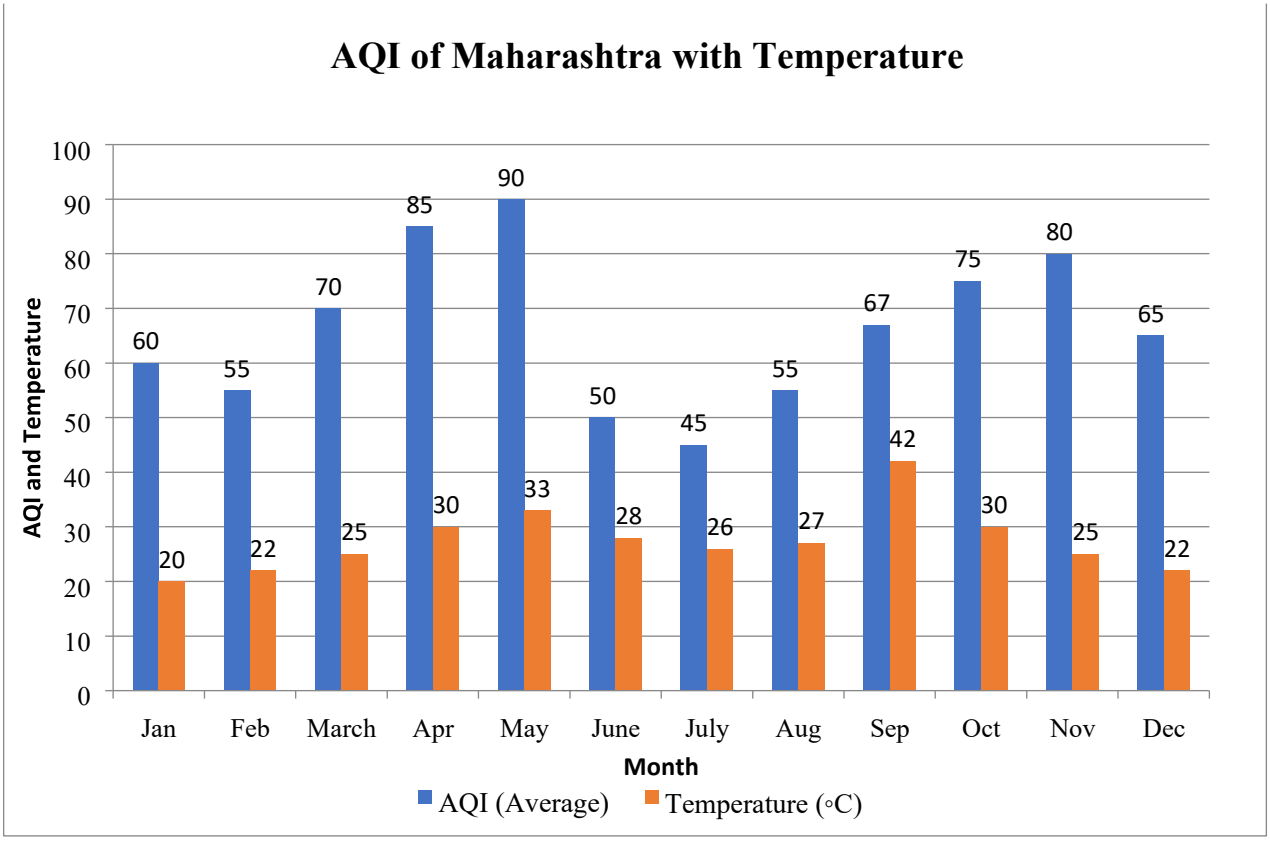


Figure 6 Depict of the AQI beside the temperature of Maharashtra for the year 2023.

Figure 6 shows the monthly average air quality index (AQI) of Maharashtra side by side with respect to temperature. The second graph shows the variations in the average AQI and temperature of Maharashtra. The AQI varies only slightly and ranges from 50-60, with a slight increase in October and November. The average temperature is approximately 22 degrees Celsius in January, approximately 32 degrees Celsius in May, and then decreases again to approximately 22 degrees Celsius in December. The parallel trend reveals an inverse relationship between the AQI and temperature during preferable temperature months.

Table 4 Statistics of various pollutants and AQIs

Pollutant	**PM2.5**	**PM10**	**NO2**	**SO2**	**O3**	**AQI**
Count	345	345	345	345	345	345
Mean	42.52	58.92	28.61	10.96	27.93	105.31
Median	42.89	58.55	28.75	11.03	28.07	105.48
Q1 (25%)	35.09	48.78	24.84	9.02	22.82	90.24
Q3 (75%)	50.08	69	32.54	12.92	33.03	120.43
10th Percentile	29.34	41.33	21.94	7.36	17.84	75.62
90th Percentile	55.91	78.62	35.67	14.75	37.21	135.89
Std Dev	9.57	14.93	5.07	2.91	6.89	19.86
Variance	91.61	222.98	25.7	8.48	47.52	394.41
Skewness	-0.02	0.05	-0.03	-0.02	0.01	-0.01

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The PM10 concentration and AQI had the greatest variability, indicating significant seasonal pollution swings. All the pollutants have a mean that is close to the median, indicating a balanced distribution with little skewness. The air quality index (AQI) routinely exceeds 100, indicating moderate to unhealthy conditions. PM2.5, NO2, SO2, and O3 levels are steady, whereas PM10 swings the greatest. The low skewness (-0.03 to 0.05) indicates that the pollution data follow a normal distribution without extreme outliers.

Table 5 Correlations between AQI and other pollutants

Feature	PM2.5	PM10	NO2	SO2	O3	AQI
PM2.5	1	0.85	0.72	0.65	-0.3	0.96
PM10	0.85	1	0.74	0.6	-0.35	0.98
NO2	0.72	0.74	1	0.55	-0.4	0.94
SO2	0.65	0.6	0.55	1	-0.25	0.88
O3	-0.3	-0.35	-0.4	-0.25	1	-0.5
AQI	0.96	0.98	0.94	0.88	-0.5	1

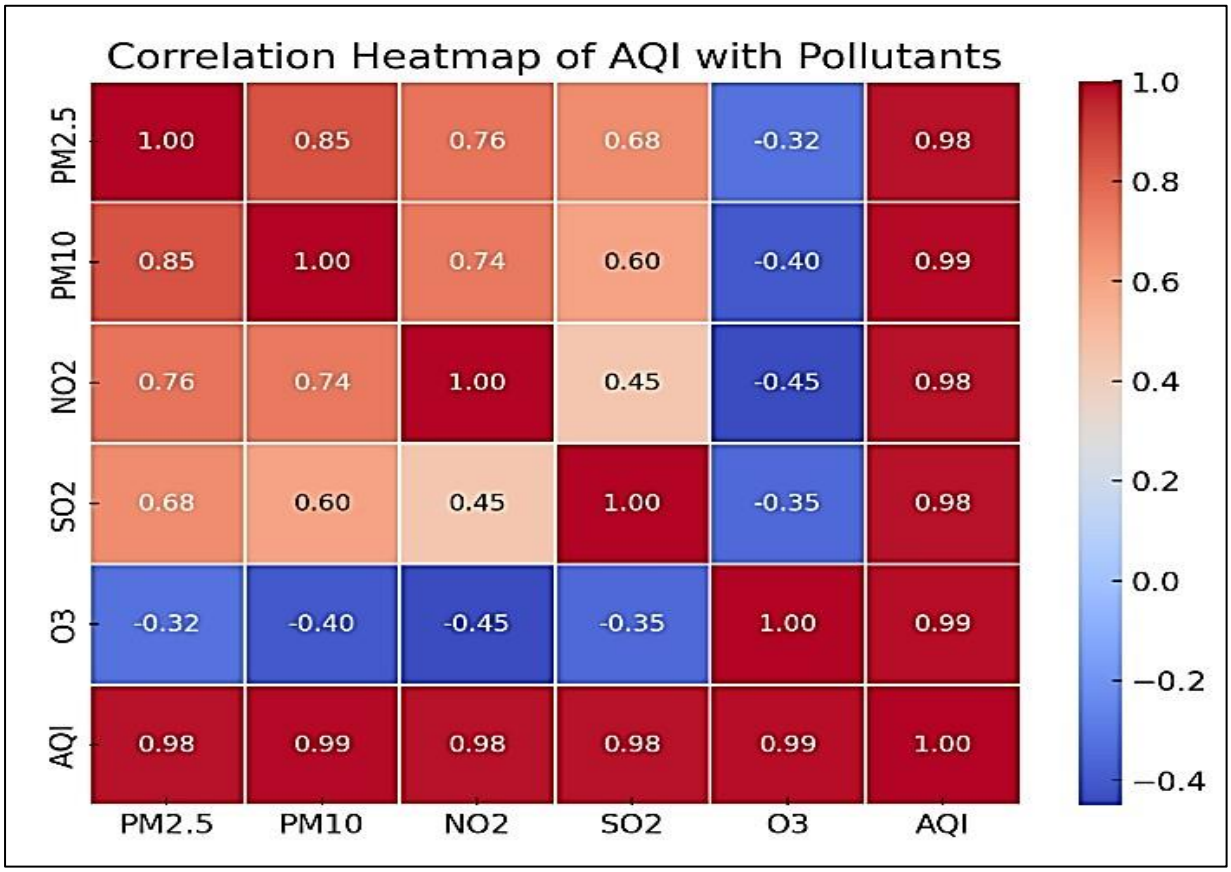


Figure 7 Heatmap correlation matrix

PM10 and PM2.5 are the primary causes of air quality degradation. NO2 and SO2 emissions from traffic and industry have a major impact on the air quality index. Ozone behaves differently, resulting in complex atmospheric chemistry at play. Reducing particulate matter and NO2 emissions can significantly improve the air quality index (AQI).

Performance Comparison

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This section highlights a comparison between the five machine learning algorithms, namely XGBoost, SVM, NBC, ANN and KNN, in terms of their ability to forecast air pollution in different cities in Maharashtra. XGBoost is expected to improve in accuracy over the other algorithms in light of the algorithm's ability to handle interactions within the data. The results of the performance comparison of various machine learning methods used in the prediction of air pollution in different cities of Maharashtra. The evaluated models include the support vector machine (SVM) (Kumar & Pande., 2023), naive Bayes classifier (NBC) (Hussain, et al., 2020), artificial neural network (ANN) (Pandya, et al., 2020), K-nearest neighbor (KNN) (Londhe & Mayuresh, 2021) and XGBoost.
Comparisons were made on the basis of accuracy, precision, F1-score, and recall. Among the models, XGBoost had a higher accuracy of 98.83%, precision of 97.70%, F1-Score of 98.14 and recall of 98.64%, respectively. The naive Bayes classifier started showing impressive results, with a high F1score of 97.78% and a high accuracy of 95.58%. Compared with the other models, it is ranked low, as SVM yields an accuracy of only 81%. In addition, XGBoost is superior in terms of air pollution forecasting for all metric types and achieves the best balance.

Table 6 Comparative analysis of ML models for predicting air pollution in Maharashtra

Model	Accuracy	Precision	F1-Score	Recall
SVM	81	90	88	93
NBC	95.58	92.25	97.78	90.24
ANN	84.79	95	90.8	96.77
KNN	83.2	92	82	94
XGBoost	98.83	97.70	98.14	98.64

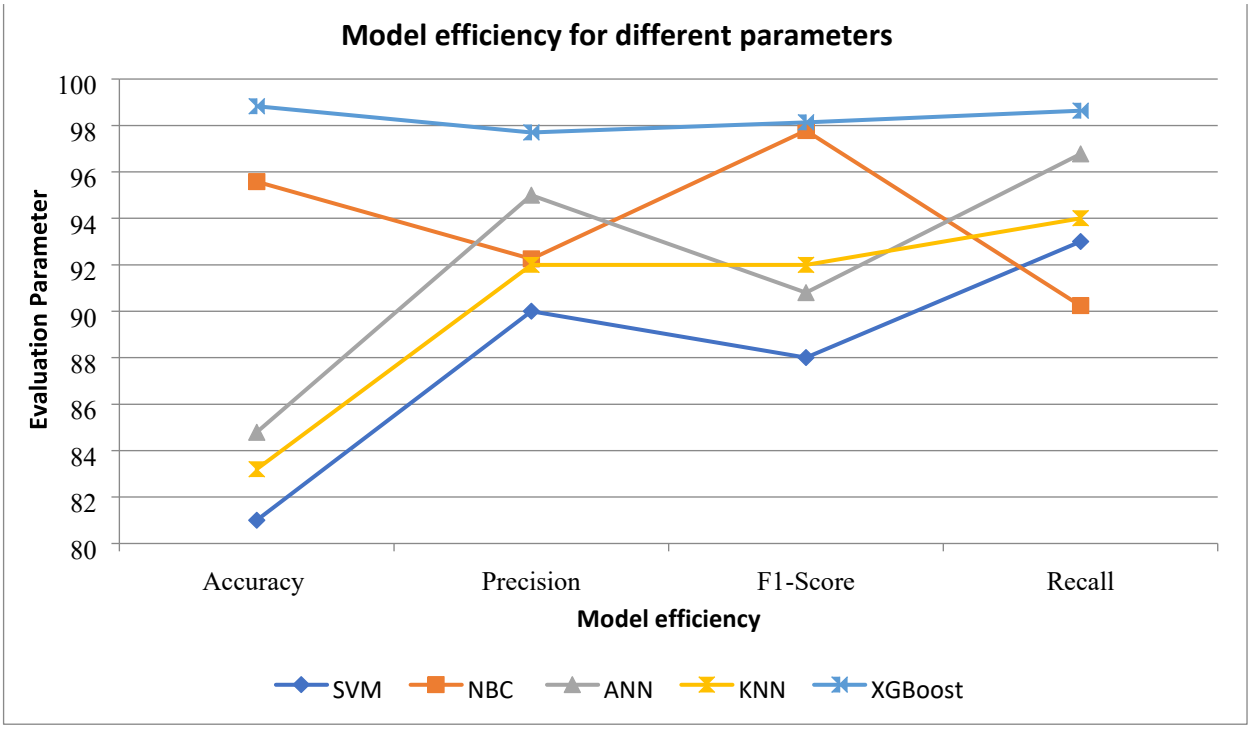


Figure 8 Visual depiction of model efficiency for different evaluation parameters.

Figure 8 visually compares the performance of the five machine learning models—SVM, NBC, ANN, KNN, and XGBoost—across four metrics: accuracy, precision, F1 score and recall. XGBoost performed better than the other methods did in each of the four-performance metrics and remained as

Received: August 06 , 2025

Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

close as possible to an idealized trend line of 100%. The NBC closely competes with XGBoost and the F1 score but has slightly lower recall values. The accuracy of the SVM model is the lowest at approximately 81 accuracy values and 93 recall values among all four models under consideration and is significantly less accurate than they are. The ANN and KNN models oscillate, and both models have good accuracy and recall values of approximately 84 yet low F1 scores, indicating model instability during prediction. In summary, XGBoost performs quite stably and retains its reliability across all the experiments.

Consequently, the findings of this study suggest a rejection of all the null hypotheses outlined in the study. The subsequent analysis involving the five different machine learning algorithms for N01 shows that XG Boost has much better performance than the other algorithms do, with an accuracy of 98.83, indicating that there is a stark difference in performance. This result provides empirical support for the H11 hypothesis that at least one algorithm has better predictive performance with respect to air pollution levels in different cities in Maharashtra. Concerning H02, the effectiveness of XGBoost means that the use of Harris hawk optimization (HHO) probably improves the choice of features, as evidenced by the improvement in the accuracy and precision indicators, and therefore confirms the second part of the alternative hypothesis H12. Finally, for H03, including the feature “land surface temperature (LST)” added to the model improved the precision and recall in support of the alternative hypothesis H13. Overall, the findings support the use of machine learning approaches, especially XGBoost, to predict air quality and therefore reject all three null hypotheses for their respective alterative hypotheses.

Conclusion

The paper satisfactorily illustrates the application of machine learning models to predict levels of air pollution alongside varying AQIs among cities in Maharashtra until 2023. Importantly, XGBOOST outperforms the other models in terms of accuracy, with a mean accuracy of 98.83%, high precision of 97.70%, high F1 score of 98.14 and high recall of 98.64%. Compared with the former, the NBC accuracy and F1 score were 95.58% and 97.78%, respectively. On the other hand, the SVM yielded the lowest accuracy (81%) of the four classifiers. Using the AQI, it was established that some of the worst affected cities, such as Mumbai and Bhiwandi, experienced increased pollution during the October and November months, which calls for timely interventions. The results could assist policymakers in realizing the applicability of ML in the identification of efficient solutions for air quality management, thereby improving health standards in Maharashtra. To further build on this work, future studies could incorporate other environmental factors into the model, including traffic density and industrial pollution. Furthermore, investigating real-time data as a concept and applying models to smart city platforms may be helpful for addressing urgent air quality issues or developing correct policies in the state of Maharashtra.

Declaration

Competing interest: The authors declare that they have no competing interest.
Data availability: The dataset will be made available upon request.

Reference

1. (n.d.). Retrieved from www.mpcb.gov.in: https://www.mpcb.gov.in/air-quality/archive/envdata
2. (n.d.). Retrieved from www.aqi.in: https://www.aqi.in/dashboard/india/maharashtra?countryfind=maharashtra+2023
3. (n.d.).
4. (2011). Retrieved from www.census2011.co.in: https://www.census2011.co.in

Received: August 06 , 2025

5. Agarwal, Aviral, Kaushik, A., Kumar, S., & Mishra., R. K. (2020). Comparative study on air quality status in Indian and Chinese cities before and during the COVID-19 lockdown period. *Air Quality, Atmosphere & Health*, 1167-1178.
6. Alade, I., Rahman, M., & Saleh, T. (2019). Modeling and prediction of the specific heat capacity of Al2 O3 water nanofluids using hybrid genetic algorithm/support vector regression model. *Nano-Structures & Nano-Objects; Elsevier*, 103-111.
7. Al-Jamimi, H., & Saleh, T. (2019). Transparent predictive modelling of catalytic hydrodesulfurization using an interval type-2 fuzzy logic. *Journal of Cleaner Production*, 1079-1088.
8. Artis, A. D., & Carnahan, H. W. (1982). Survey of emissivity variability in thermography of urban area. *Remote Sensing of Environment*, 313–329.
9. Ayturan, Y., Ayturan, Z., Altun, H., C, K., Tuncez, F., Dursun, S., & Ozturk, A. (2020). Short-term prediction of PM2.5 pollution with deep learning methods. Global NEST J. 126–131.
10. Butler, M. T., Lawrence, G. M., Gurjar, R. B., Aardenne, J. V., Schultz, M., & Lelieveld, J. (2008). The representation of emissions from megacities in global emission inventories. *Atmospheric Environment, 42*, 703–719.
11. Chauhana, V. S., Singha, B., Ganesha, S., Chauhanb, D. S., Gupta, S., Sharma, G., & Zaidib, J. (2020).
12. AIR POLLUTION IN JHANSI: AIR QUALITY INDEXING AND STATISTICAL DATA ANALYSIS. 1167-1178.
13. Das, A., & Ghosh, A. (2023). Landscape assessment of the cities in the state of Maharashtra: First step towards air quality management (AQM) and strategic implementation of mitigation plans. *Environmental Science and Pollution Research,* 59233-59248.
14. Heidari, A. A., iMirjalili, S., Hossam, F., Ibrahim, A., Majdi, M., & Huiling, C. (2019). Harris hawk's optimization: Algorithm and applications. *Future Generation Computer Systems; Elsevier*, 849872.
15. Hussin, M., & Sulaiman, M. N. (2015). A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. *International Journal of Data Mining & Knowledge Management Process*.
16. Hussain, A., Draz, U., Ali, T., Tariq, S., Irfan, M., Glowacz, A., . . . Saifur, R. (2020). Waste Management and Prediction of Air Pollutants Using IoT and Machine Learning Approach. *Energies*.
17. J., E. (1994). GIS-based approach to microclimate monitoring in Singapore's high-rise housing estates. *Photogrammetric Engineering and Remote Sensing*, 1225–1232.
18. Khedekar, S., & Thakare, S. (2023). Correlation analysis of atmospheric pollutants and meteorological factors using statistical tools in Pune, Maharashtra. *In E3S Web of Conferences. 391*, p. 01190. EDP Sciences.
19. Kulkarni, G. E., Muley, A. A., Deshmukh, N. K., & Bhalchandra, P. U. (2018). Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India. *Modeling Earth Systems and Environment, 4*, 1435-1444.
20. Kulkarni, P. S., & Jadhav, D. O. (2025). An optimization framework forelectronic waste recycling:
21. integrating cost-effectiveness and environmental considerations through binary integer programming. *Environmental Monitoring and Assessments*. Doi: https://doi.org/10.1007/s10661025-14108-0
22. Kumar, K., & Pande., B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, 5333-5348.

23. Kumar, K., Chaudhri, S. N., Rajput, N. S., Shvetsov, V. A., Sahal, R., & Alsamhi, S. H. (2023). An IoT-enabled E-Nose for remote detection and monitoring of airborne pollution hazards using LoRa network protocol. *Sensors*, 4885.
24. Kumari, M., Somvanshi, S. S., & Zubair, S. (2021). Estimation of air pollution using regression modelling approach for Mumbai region, Maharashtra, India. *Remote Sensing and GI Science:*
25. *Challenges and Future Directions*, 229-247.
26. *Landsat 7 Data Users Handbook.* (2002).
27. Londhe, & Mayuresh. (2021). Data mining and machine learning approach for air quality index prediction. *International Journal of Engineering and Applied Physics*, 136-153.
28. Mangalekar, S. B., Jadhav, A. S., & Raut, P. D. (2015). Studies on ambient air quality status of Kolhapur city, Maharashtra, India during year 2013. *Asian Journal of Water, Environment and Pollution*, 15-22.
29. Molina, J. M., & Molina, L. T. (2004). Megacities and atmospheric pollution. *Journal of the Air & Waste Management Association*, 644–680.
30. N. K. Rai, T. S., & Vyas, A. (2017). Air quality index determination of residential areas of Jodhpur city. *International Journal of Current Advanced Research*, 7046-7048.
31. NYT. (2014). India's Air Pollution Emergency [online]. *Current World Environment 18*.
32. Onkal-Engin, G., Demir, I., & Hiz, H. (2004). Assessment of Urban air quality in industrial using fuzzy synthetic evaluation. *Atmospheric Environment, 38*, 3809-3815.
33. Pandya, S., Ghayvat, H., Sur, A., Awais, M., Kotecha, K., Saxena, S., . . . Pingale, G. (2020). Pollution weather prediction system: smart outdoor pollution monitoring and prediction for healthy breathing and living. *Sensors*, 5448.
34. Pathan, M. H. (2022). AMBIENT AIR QUALITY MONITORING IN A POPULAR TOURIST DESTINATION IN MAHARASHTRA AURANGABAD. *International Journal of Advance and Applied Research*.
35. Patil, M., Radhika, & Dinde, H. T. (2023). Status of ambient air pollution in different states of India during 1990-2015. *Current world Environment*, 245.
36. Priya, R. M., & Sathya, P. (2019). statistical analysis of air pollutants in ambient air, reality of sensors and corrective measures in India. *Innovations in Power and Advanced Computing Technologies, 1*(2019), 1-6.
37. Purohit, A., Chauhan, P., Vyas, M., Vyas, D. A., & Sing, D. S. (2017). AIR QUALITY INDEX DETERMINATION OF COMMERCIAL AREAS OF JODHPUR CITY: A Case Study. *4*, 4.
38. Q. Weng, & S. Yang. (2006). Urban air pollution patterns, land use and thermal landscape: An examination of the linkage using GIS. Environmental Monitoring and Assessment. 463–489.
39. Rai, N. K., & Sharma, T. A. (n.d.). Air quality index determination of residential areas of jodhpur city: a case study. 3.
40. Rao, M. N., & N., H. V. (2001). *Air pollution.* New Delhi: Tata McGraw Hill Publishing Company Limited.
41. Rao, M. P., Hima, B. V., G. S., Indracanti, J., & Anjaneyulu, Y. (2004). Assessment of ambient air quality in the rapidly industrially growing Hyderabad urban environment. *Workshop program and presentation.* Proceedings of the BAQ.
42. Rathi, D. B., & Rathod, S. D. (2016). Statistical Analysis of Ambient Air Quality in Aurangabad. *Mod Chem appl 4*.
43. S, M. (2006). Chemical evolution of gaseous air pollutants down-wind of tropical megacities: Mexico City case study. *Atmospheric Environment*, 6012–6018.
44. Schwela, & Dietrich. (2000). Air pollution and health in urban areas. *Reviews on environmental health 15, no. 1-2*, 13-42.
45. Shama, N., Taneja, S., & Bhatt, A. (2020). Empirical Analysis of life quality based on air pollution in states of India. *Journal of Statistics and Management system*, 1213-1226.

Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

46. Sharma, G. M., & Kota, S. H. (2024). Mapping air quality trends across 336 cities in India: Insights from three decades of monitoring (1987–2019). *Environment International*, 108979.
47. Snyder, W. C., Wan, Z., Zhang, Y., & Feng, Y. Z. (1998). Classification-Based Emissivity for Land Surface Temperature Measurement from Space. *International Journal of Remote Sensing*, 2753-2774.
48. Sonar, D. C., & TAYADE, D. A. (2015). Statistical Analysis of Major Parameters of Ambient Air Quality. *International Journal of Modern Sciences and Engineering Technology (IJMSET), 2*(8), 13-21.
49. Sonavane, P., & Pinjar, S. (2024). Assessment of the Ambient Air Quality of Sangli City, Maharashtra, vis-à-vis National Ambient Air Quality Standards (NAAQS). *International Research Journal on Advanced Engineering Hub (IRJAEH), 2*, 182-189.
50. TAYADE, A. Y. (2012). Statistical Analysis of Ambient Air Quality in Mumbai City by Using Air Quality Index and Seasonal Variations. *International Journal of MATHEMATICS AND APPLIED STATISTICS, 3*, 95-102.
51. Vashisht, A., Somvanshi, S., & Shrivastava, P. (2018). Spatiotemporal modelling of air quality of Delhi using remote sensing and GIS. *National Conference on Environmental challenges for New Delhi.* ESDA.
52. WB. (2013;2015, January 23). Diagnostic assessment of select environmental challenges in India. *The World Bank*.
53. Yan, Z., Hao Chen, X. D., & Zhigang, X. (2022). Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost. *Expert Systems with Applications*, 117943.
54. Zhang, K., Chen, Y., & Wu, L. (2019). Grey Spectrum Analysis of Air Quality Index and Housing Price in Handan . *Complexity, 2019*, 1-6.
55. Figure 1 A schematic diagram that illustrates primary aspects in evaluating the exposure to air pollution

Figure 2 Flow chart of data cleaning 8
Figure 3 Fluctuations in AQI levels among various cities in Maharashtra..... 11
Figure 4 Depiction of AQI of different cities of Maharashtra 12
Figure 5 Monthly patterns of air contaminants (PM2.5, PM10, Ozone and SO2) and temperature in Maharashtra..... 13
Figure 6 Depict the AQI beside the temperature of Maharashtra for the year 2023. 14
Figure 7 Heatmap correlation Matrix 16
Figure 8 Visual depiction of model efficiency based of different evaluation parameter. 17

Table 1 AQI scale for rating..... 2546
Table 2 AQI of several cities in Maharashtra for the year 2023 2548
Table 3 Monthly Average AQI and Key Pollutant Levels in Maharashtra for 2023 2550
Table 4 Statistics of various pollutants and AQI..... 2552
Table 5 Correlation between AQI and other pollutants 2553
Table 6 Comparative analysis of ML models for predicting air pollution in Maharashtra..... 2554

Received: August 06 , 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Received: August 06 , 2025