PLOS One

RESEARCH ARTICLE

Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM

Sevtap Tırınk*

Department of Medical Services and Techniques, Environmental Health Program, Iğdır University, Iğdır, Türkiye

* sevtap.tirink@igdir.edu.tr

Abstract

Air pollution is a global problem that threatens environmental sustainability and severely affects public health. Monitoring air quality and predicting future pollution levels are critical for creating effective environmental policies and enabling individuals to take precautions against air pollution. This study presents a long-term assessment of daily Air Quality Index (AQI) prediction using machine learning models based on meteorological and pollutant data collected in eastern Türkiye from 2016 to 2024. The dataset includes four major air pollutants (PM_{10}, SO_2, NO_2, O_3) and five meteorological variables (temperature, precipitation, relative humidity, wind direction, wind speed). Three models—eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine (SVM)—were evaluated using the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) as performance metrics. Among these, XGBoost achieved the highest prediction accuracy ($R^2 = 0.999$, RMSE = 0.234, MAE = 0.158). The results demonstrate that ensemble-based machine learning approaches, particularly XGBoost, can effectively model AQI fluctuations using environmental predictors. These results provide valuable insights for air quality forecasting systems and suggest practical implications for regional air pollution management and early warning systems, supporting public health protection and the development of environmental health policies.

Citation: Tırınk S (2025) Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM. PLoS One 20(10): e0334252. https://doi.org/10.1371/journal.pone.0334252

Editor: Takayuki Mizuno, National Institute of Informatics, JAPAN

Received: August 2, 2025

Accepted: September 22, 2025

Published: October 8, 2025

Data availability statement: The data underlying the results presented in this study were obtained from two official databases of the Ministry of Environment, Urbanization and Climate Change (Türkiye). Air quality monitoring data are publicly accessible at

Introduction

Air is an indispensable resource for the sustainability of life, and its quality is critical to human health and the environment. However, air pollution has become a serious global environmental problem that negatively affects atmospheric quality and public health [1,2]. Rapid industrialization and urbanization have significantly increased the

https://sim.csb.gov.tr/. Meteorological data are available through the General Directorate of Meteorology's MEVBIS system (https://mevbis.mgm.gov.tr/) and require user registration and purchase via the Ministry's official platform. Additional information regarding access procedures can be obtained by contacting the General Directorate of Meteorology at pazarlama@mgm.gov.tr. These data are available to other researchers under the same conditions as those applied to the author.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: RMSE, Root Mean Square Error; MLR, Multiple Linear Regression; MSE, Mean Squared Error; MAE, Mean Absolute Error; R^2, Coefficient Of Determination; SD, Standard Deviation; T, Temperature; P, Precipitation; RH, Relative Humidity; WD, Wind Direction; WS, Wind Speed; RF, Random Forest; RFR, Random Forest Regression; CR, Catboost Regression; DT, Decision Tree; MP, Multilayer Perceptron; GB, Gradient Boosting; LR LinearR, Lasso Regression; SM, Stacked Models; KNN, K-Nearest Neighbor; LightGBM, Light Gradient Boosting Machine; SVR, Support Vector Regression; LSTM, Long Short-Term Memory; LogR, Logistic Regression; MLR, Multiple Linear Regressive; ANN, Artificial Neural Networks; ILSTM, Improved Long Short-Term Memory Algorithm; BRNN, Bayesian Regularized Neural Networks.

concentration of harmful substances in the atmosphere, especially in industrial areas, resulting in increased pollutant emissions [3,4]. This trend has increased public awareness and concern about air quality, as well as triggered calls for the development of effective air quality management and pollution control strategies [5].

The urbanization process plays a decisive role in air pollution levels. The spatial arrangement of urban areas affects the distribution and concentration of pollutants, causing denser urban environments to be exposed to higher pollution levels, usually due to traffic emissions and industrial activities [4,5]. In the Iğdır province of Türkiye, rapid economic expansion and urbanization processes also lead to a severe deterioration in air quality. In many regions of Türkiye, especially in the winter months, using fossil fuels for heating is one of the main factors increasing air pollution. In regions such as Iğdır, this situation becomes more pronounced in the winter months, and air quality deteriorates significantly [6,7].

The effects of air pollution on public health are profound. Various studies have shown that exposure to major air pollutants — including particulate matter ($PM_{2.5}$ and PM_{10}), sulfur dioxide (SO_2), nitrogen oxides (NOx), carbon monoxide (CO), and ozone (O_3) — is associated with serious health problems such as respiratory diseases, cardiovascular diseases, and even cancer [8,9]. In this regard, Kumar et al. [10] highlighted how particulate matter, together with meteorological conditions, critically contributes to poor air quality and adverse health outcomes. The world health organization emphasizes the urgent need for comprehensive monitoring and management strategies because of premature deaths from air pollution [11]. Increasing urban populations makes air quality protection more critical, necessitating addressing both local and transboundary pollution sources [12]. In this context, authorities are trying to combat air pollution by developing long-term strategies. However, since these solutions require large-scale implementation, they can be costly in terms of time and financial resources. Providing air quality estimates and air quality index (AQI) values is important for individuals to develop protection strategies [13]. Air quality is influenced by the interactions among economic development, urban planning, and public health. Forecasting air pollution can support local governments and vulnerable groups (e.g., individuals with respiratory diseases, pregnant women, and children) by indicating when and where air quality may deteriorate and enabling timely protective measures.

As a widely used index, the AQI provides an overall assessment of environmental air quality. It expresses the overall air quality with a single numerical value by combining the concentrations of specific pollutants (Particulate Matter smaller than 2.5 micrometers ($PM_{2.5}$), Particulate Matter smaller than 10 micrometers (PM_{10}), Ozone (O_3), Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), and Sulfur Dioxide (SO_2)) [14,15]. The Turkish National Air Quality Index (TNAQI) is the version of AQI developed by the US Environmental Protection Agency (EPA) and adapted to national legislation and limit values [16]. On a scale from 0 to 500, the AQI is divided into six categories – good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, and hazardous – with corresponding health warnings provided in Table 1 [16]. As the AQI value increases, the threat to human health also increases.

Table 1. Value ranges of AQI and corresponding health warnings.

Values of index	Levels of concern	AQI colour	Description of air quality
0–50	Good	Green	Air quality is satisfactory and air pollution poses little or no risk.
51–100	Moderate	Yellow	Air quality is acceptable, but some pollutants may pose a moderate health concern for a small number of people who are unusually sensitive to air pollution.
101–150	Unhealthy for sensitive groups	Orange	Health impacts may occur for sensitive groups. The general public is unlikely to be affected.
151–200	Unhealthy	Red	Everyone may begin to experience health effects, with serious health effects for sensitive groups.
201–300	Very unhealthy	Purple	May constitute a health emergency. The entire population is likely to be affected.
301–500	Hazardous	Maroon	Health alert: Everyone may experience more serious health effects.

https://doi.org/10.1371/journal.pone.0334252.t001

The relationship between air quality and meteorological factors is shaped by complex chemical and dynamic atmospheric reactions. Concentrations of air pollutants are highly sensitive to meteorological conditions such as wind speed and direction, relative humidity, and temperature, which have been shown in various studies to directly affect local air pollution levels [7–17]. Sekula et al. [18] evaluated the effect of atmospheric circulation on air quality, emphasizing the importance of humidity and temperature gradients, especially in the lower troposphere. Therefore, meteorological conditions should be taken into account in air quality assessments and forecasting, since they critically influence pollutant dispersion rather than being controllable factors for improving air quality.

In recent years, the use of machine learning algorithms in air quality prediction has become increasingly widespread, and significant progress has been made in this field. Studies conducted in this field in recent years are summarized in Table 2, where model performances are primarily assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The table includes a variety of machine learning methods such as Random Forest (RF), Random Forest Regression (RFR), Cubist Regression (CR), Decision Tree (DT), Multilayer Perceptron (MP), Gradient Boosting (GB), Linear Regression (LR), Stacked Models (SM), k-Nearest Neighbor (KNN), Light Gradient Boosting Machine (LightGBM), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), Logistic Regression (LogR), Artificial Neural Network (ANN), Improved Long Short-Term Memory (ILSTM), and Bayesian Regularized Neural Networks (BRNN). Meteorological variables are abbreviated as Temperature (T), Precipitation (P), Relative Humidity (RH), Wind Direction (WD), and Wind Speed (WS). This progress results from collaborative efforts, as various studies examine the methods developed to increase the effectiveness of machine learning techniques in air quality prediction [19–21].

As shown in Table 2, ensemble learning methods such as XGBoost, and LightGBM frequently achieved the highest accuracy in AQI prediction across different regions. Support vector-based models also provided competitive results in several studies, particularly for PM-based predictions. In most cases, particulate matter ($PM_{2.5}$ and PM_{10}) emerged as the dominant factor influencing air quality, highlighting its critical role in AQI forecasting. These studies demonstrate the growing interest in AQI prediction using machine learning. However, most of them focus on short-term datasets, metropolitan regions, or limited sets of pollutants and meteorological variables. To the best of our knowledge, no long-term, region-specific study has been conducted in a geopolitically sensitive area like Iğdır, Türkiye. In addition, existing models rarely integrate both pollutant and meteorological data over extended periods. This gap further highlights the novelty of the present research.

Air pollution prediction varies across different regions and climatic conditions due to the complex, nonlinear interactions between atmospheric pollutants and meteorological parameters, posing a significant challenge on a global scale [34–36]. Although machine learning models are widely used for AQI prediction, comprehensive evaluations of the forecast

Table 2. Summary of studies applying machine learning models to predict the AQI.

Method	Result(s) for best model	Regions	Factors	Author(s) and Year
Adaboost, Catboost, GB, KNN, Linear Regression, RF, SVM, XGBoost	For AQI: $R^2=0.9850$, RMSE = 11.2696, MAE = 8.3845 (XGBoost)	Azamgarh, India	$PM_{2.5}$, PM_{10}, NO_2, SO_2, T, Humidity, WD, WS, UV Radiation	[22]
KNN, SVM, DT, RF, MP, GB, LR, SM	For AQI: $R^2=0.973$, RMSE=7.568, MAE=4.596 (SM)	Beijing, China	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	[23]
LightGBM, RF, CatBoost, AdaBoost, XGBoost	For AQI: $R^2=0.9998$, RMSE = 0.76, MAE = 0.60 (CatBoost)	Visakhapatnam, India	$PM_{2.5}$, PM_{10}, NO_x, NH_3, SO_x, CO, O_3, Benzene, Toluene, Xylene, and meteorological factors	[24]
SVR, RFR, CR	For AQI: RMSE = 0.2792, Accuracy = 79.8622% of New Delhi (CR), RMSE = 0.5674 Accuracy = 68.6860% of Bangalore (CR), RMSE = 0.0988, Accuracy = 93.7438% of Kolkata (RFR), RMSE = 0.0628, Accuracy = 97.6080% of Hyderabad (RFR)	New Delhi, Bangalore, Kolkata, Hyderabad (India)	$PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, Toluene	[25]
SARIMA, SVM (RBF Kernel), LSTM	For AQI: $R^2=0.9989$, RMSE = 4.944 (SVM with RBF kernel)	Ahmedabad city of Gujarat, India	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3, NH_3, Pb	[26]
LR, KNN, SVR, LSTM, RF, XGBT, LightGBM, LSTM-SVR (Hybrid)	For AQI: $R^2=0.994$, RMSE = 3.277, MAE = 1.754 (LSTM-SVR)	Six major Chinese urban agglomerations	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3, T, Pressure, Dew Temperature, WS, P	[27]
DT, RF, XGBoost	For AQI: $R^2=0.9214$, RMSE = 29.6953, MAE = 18.9839 (XGBoost)	India	Xylene, PM_{10}, NH_3, Toluene, Benzene, $PM_{2.5}$, NO_x, O_3, SO_2, NO_2, NO, CO	[13]
LR, KNN, SVM, LSTM, GRU, Hybrid LSTM-GRU	For AQI: MAE = 36.11, RMSE = 52.03, $R^2=0.84$ (Hybrid LSTM-GRU)	Delhi, India	$PM_{2.5}$, NO_x, O_3, SO_2, NO_2, NO, CO, T, WS, RH, Solar Radiation	[28]
DT, LogR, RF	For AQI: Accuracy = 98.63% (DT)	Uttarakhand state, India	PM_{10}, $PM_{2.5}$, SO_2, NO_2	[29]
ANN, MLR, ILSTM, SVR	For AQI: $R^2=0.981$ (MLR)	Tamil Nadu State, India	$PM_{2.5}$, PM_{10}, SO_2, NO_x, NH_3, CO and O_3	[30]
SVR and LSTM	For AQI: RMSE = 10.995, $R^2=0.570$ (LSTM)	Chennai city, India	$PM_{2.5}$, NO_2, SO_2, CO, Ozon	[31]
SVR-RBF and PCA SVR-RBF	For AQI: accuracy = 94.1% (SVR-RBF)	California	CO, NO_2, SO_2, Ozone, $PM_{2.5}$, WS, T, RH	[32]
SVR and RFR	For AQI: $R^2=0.9766$ RMSE = 7.666 (SVR)	Beijing and Italian	$PM_{2.5,}$ PM_{10}, O_3, SO_2, NO_2	[33]

https://doi.org/10.1371/journal.pone.0334252.t002

performance of these models are often limited to short-term datasets or specific subsets of pollutants [37–39]. This study aims to (i) reveal the long-term status and temporal dynamics of air pollution in Iğdır, Türkiye, during 2016–2024 using the TNAQI together with pollutant and meteorological records and (ii) systematically compare the predictive performances of advanced machine learning models (XGBoost, LightGBM and SVM) for daily AQI prediction and quantitatively evaluate the impacts of air pollutants and meteorological factors. Unlike traditional studies that usually focus on short-term datasets or specific subsets of pollutants, this research presents a comprehensive evaluation by integrating multiple air pollutants (PM_{10}, SO_2, NO_2, O_3) with key meteorological variables to improve the forecast accuracy. The findings provide valuable insights into the adaptability of machine learning models under different environmental conditions, suggesting a scalable and generalizable AQI prediction framework that can contribute to developing data-driven air quality management policies globally.

The study's objectives are based on these observations:

a) Calculate daily AQI using data from four air pollutants (PM_{10}, SO_2, NO_2, O_3) and five meteorological parameters (temperature, precipitation, relative humidity, wind direction, wind speed) from 2016 to 2024.

b) Apply and compare advanced machine learning models, including XGBoost, LightGBM, and SVM, to predict the AQI with high precision.

c) Evaluate the performance of the machine learning models using metrics such as R^2, RMSE and MAE and assess their predictive capabilities.

d) Train the models on 80% of the dataset and validate them using the remaining 20% to ensure reliability and accuracy in AQI prediction.

e) Perform a comparative analysis to identify the most suitable model for AQI prediction, highlighting XGBoost's consistent performance across all metrics.

f) Analyze model results to determine influential parameters impacting AQI predictions and offer insights for efficient air quality management.

This study offers a novel contribution to the field of air quality modeling by focusing on Iğdır province, a border region of Türkiye adjacent to three countries—Iran, Armenia, and Nakhchivan (Azerbaijan). Despite its strategic geopolitical location and vulnerability to cross-border pollution transport, the region has remained underrepresented in empirical AQI forecasting studies. By integrating eight years of daily meteorological and pollutant data (2016–2024), this study systematically compares the predictive performance of three advanced machine learning models—XGBoost, LightGBM, and SVM. The findings demonstrate that XGBoost significantly outperforms the other models in terms of accuracy and generalizability. This comprehensive regional analysis not only fills a critical gap in the existing literature but also provides a scalable framework for AQI prediction in other transboundary or climatically similar regions.

Materials and methods

Study area and data description

The current study examined the AQI of Iğdır City of Türkiye. Iğdır province is a settlement located east of Türkiye and borders three countries (Iran, Armenia, and Nakhchivan). The coordinates of Iğdır province are recorded as latitude: 39° 55′ 25.36″ N and longitude: 44° 02′ 42.00″ E. Its surface area is 3588 km^2. Also, Iğdır, located at an altitude of 800–900 m, has a provincial population of approximately 210000. The geographical location of the study area is illustrated in Fig 1.

Air quality and meteorological data, including PM_{10}, NO_2, SO_2, O_3, wind speed, wind direction, relative humidity, temperature, and precipitation, were collected from air quality and meteorology stations operated by the Republic of Türkiye Ministry of Environment, Urbanization and Climate Change. The dataset comprises eight year daily air pollutants data (2016–2024) for Iğdır.

Methodology for AQI calculation and indexing

In this study, AQI values were calculated using pollutant concentrations measured at the air monitoring station in Iğdır province. The TNAQI system, adapted to Türkiye's national standards, was applied. According to this approach, the AQI is determined as the maximum sub-index among five pollutants (SO_2, NO_2, PM_{10}, O_3, CO). Reference intervals for pollutant concentrations and their corresponding AQI values are provided in Table 3 [16].

The AQI calculation follows Equations (1) and (2). For this purpose, four pollutants (PM_{10}, SO_2, NO_2, O_3) measured in Iğdır were used:

$$AQI = \max\left(IAQI_1,\ IAQI_2,\ \ldots,\ IAQI_i\right) \tag{1}$$

$$IAQI_i = I_{low} + \frac{C_i - C_{low}}{C_{high} - C_{low}}\left(I_{high} - I_{low}\right) \tag{2}$$

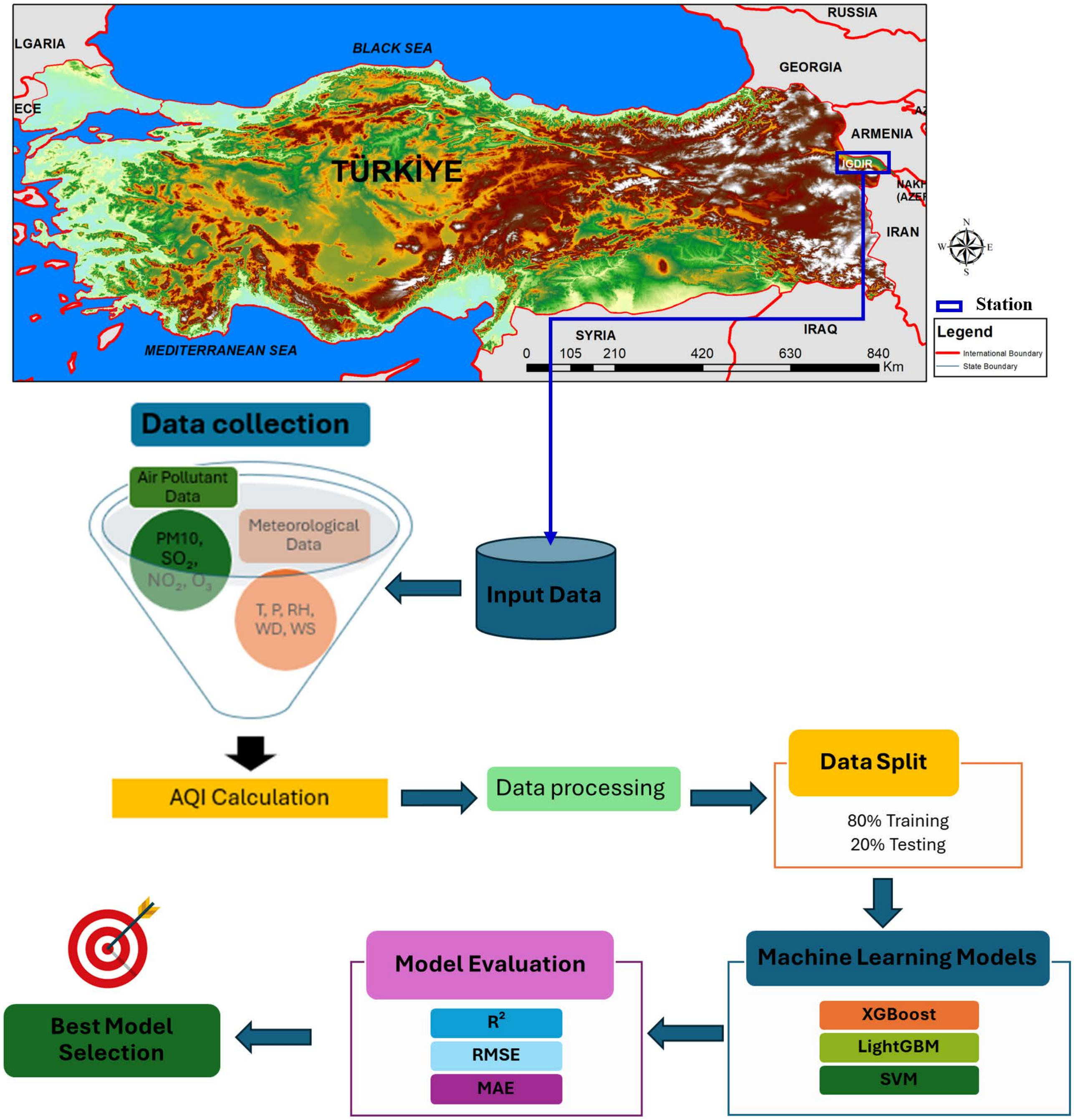

Fig 1. Location map of the study area and flowchart of the methodology. The study area is located in Iğdır Province, Türkiye. The base map is sourced from Natural Earth (public domain; http://www.naturalearthdata.com) and elevation data are from the USGS National Map Viewer (public domain; https://www.usgs.gov).

https://doi.org/10.1371/journal.pone.0334252.g001

where i represents the air pollutant. $IAQI_i$ is the air quality sub-index for pollutant i; C_i is the concentration of pollutant i; C_{low} and C_{high} denote the minimum and maximum concentration values of the AQI category corresponding to the specific pollutant; I_{low} and I_{high} denote the minimum and maximum AQI values for that category (see Table 3). The following

Table 3. Reference intervals used in AQI calculation.

AQI scale	Index value	SO_2 [µg m^{-3}] 1 h. Avg.	NO_2 [µg m^{-3}] 1 h. Avg.	CO [µg m^{-3}] 8 h. Avg.	O_3 [µg m^{-3}] 8 h. Avg.	PM_{10} [µg m^{-3}] 24 h. Avg.
Good	0–50	0-100	0-100	0-5500	0-120	0-50
Moderate	51–100	101-250	101-200	5501-10000	121-160	51-100
Unhealthy for sensitive groups	101–150	251-500	201-500	10001-16000	161-180	101-260
Unhealthy	151–200	501-850	501-1000	16001-24000	181-240	261-400
Very unhealthy	201–300	851-1100	1001-2000	24001-32000	241-700	401-520
Hazardous	301–500	>1101	>2001	>32001	>701	>521

https://doi.org/10.1371/journal.pone.0334252.t003

calculations were used in the study: maximum one-hour average values (µg m^{-3}) for NO_2 and SO_2, maximum eight-hour average values (µg m^{-3}) for O_3, and daily average values for PM_{10} (µg m^{-3}). These daily AQI values, along with the daily average pollutant concentrations and meteorological parameters, were subsequently used as input for the machine learning models. The overall methodology is illustrated in Fig 1.

Machine learning models

Light gradient boosting machine algorithm (LightGBM). The LightGBM algorithm is a framework offering a gradient boosting method using tree-based learning methods [40]. It was developed by Microsoft researchers in 2017 [41]. This algorithm has a distributed structure and provides a faster training process and higher performance compared to other algorithms. LightGBM is based on a leaf-based growth strategy using one-sided gradient sampling, special feature grouping, and depth-limited histograms. This algorithm aims to achieve high prediction success by creating a strong learner from the combination of weak learners. In particular, LightGBM uses maximum depth-limited leaf-based growth and histogram-based methods to shorten the training time and reduce memory usage [42]. In addition, the histogram subtraction technique enables the use of a large number of histograms by dividing continuous explanatory variables, thus increasing statistical efficiency and accelerating convergence [43]. The LightGBM algorithm is a gradient boosting model built using tree-based classifiers [27,44,45]. Trees are constructed iteratively such that each step minimizes the loss function. However, traditional approaches often struggle with speed and capacity. To address these challenges, LightGBM efficiently handles large datasets and categorical features using techniques such as histogram-based splitting and leaf-wise tree growth [46].

eXtreme gradient boosting algorithm (XGBoost). The XGBoost algorithm is a gradient boosting method proposed by Chen and Guestrin in 2016 [47]. The XGBoost aims to make more accurate predictions by adding a regularization term to the objective function to prevent overfitting [48]. The XGBoost algorithm constructs an ensemble containing a set of decision trees trained on different dataset partitions [49]. When splitting trees by depth or level, XGBoost determines the best branch splitting effect of each tree (decision) feature and the appropriate threshold for this feature. The XGBoost performs successive splits to make the tree structures more distinct [48]. Finally, the scores of the stable trees obtained during the training process are summed, and the final predicted value of the response variable is calculated [50].

Support vector machine algorithm (SVM). The SVR algorithm is an important subgroup of support vector machine, one of the machine learning algorithms [51]. While the support vector machine algorithm used for classification operations is called support vector classification, the part that deals with modeling and prediction operations is called SVR [52,53]. Since SVR is a supervised learning method, the success of the predictions made with SVR varies depending on the training and test data sets [54]. The main goal of SVR in the linear SVR model is to define a function f(x) with the maximum deviation (ε) value from the training set and is as flat as possible. The training data lies within the limits between −ε and +ε [51]. However, many studies cannot be modeled within the framework of linear features. Therefore, in nonlinear

SVR cases, the input data is transformed into a higher dimensional hilbert space so the regression line can be linear [54]. There are many nonlinear kernel functions, including the Gaussian radial basis function kernel. The kernel function used in this study is the Gaussian radial basis function kernel.

Model comparison criteria. In this study, the performance of the models was assessed using three criteria: RMSE, MAE, and R².

Root-mean-square error. The RMSE is calculated as the square root of the Mean Squared Error (MSE), which represents the average squared difference between the observed actual output values and the model's predicted values. RMSE provides a measure of how much the predictions deviate from the actual values, effectively serving as the standard deviation of these differences. It ranges from 0 to positive infinity, where a lower RMSE indicates predictions that are closer to the target values. In comparison, a higher RMSE reflects a more significant deviation and a wider distribution of values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{3}$$

Mean Absolute Error. The MAE quantifies the error in model predictions by calculating the mean of the absolute differences between the actual and predicted values. It is obtained by summing the absolute differences for all observations in the dataset and dividing this sum by the total number of observations. The MAE ranges from 0 to positive infinity, where smaller values closer to 0 indicate better performance, as they reflect that the predicted values are more closely aligned with the actual target values.

$$MAE = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{n} \tag{4}$$

R-Squared. The R² represents how well the model predicts the target variable. This criterion ranges from 0 to 1. When R² equals 1, the predictions perfectly align with the data, indicating absolute accuracy, while lower R² values suggest weaker predictive performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}_i - y_i\right)^2} \tag{5}$$

In the above equations, ŷ represents the predicted AQI variable, y denotes the actual AQI variable, $\bar{y}$ symbolizes the average value of AQI, and i refers to the index of each data point in the dataset. The upper bounds of MAE and RMSE extend to positive infinity, which makes them less effective in fully reflecting a model's overall performance across different datasets. In contrast, R², with its range confined between 0 and 1, provides a more reliable metric for comparing model performance on varying datasets. A model is considered effective when its RMSE and MAE values are low, and its R² value approaches 1. Furthermore, while all three metrics are suitable for evaluating a model's performance on the same dataset, R² is particularly useful for comparing models across different datasets.

Software information

The dataset was randomly split into 80% training and 20% testing. 10-fold cross-validation was used for all algorithms, and although different k values were tested, the most robust results were obtained with 10-fold validation. All statistical analyses were performed using R software [55]. The "psych" package was used for descriptive statistics, and the "corrplot" package was used to visualize the relationships between explanatory and response variables, and the "caret" package was used to separate the dataset into training and test sets [56–58]. For the implementation of LightGBM, XGBoost, and SVR algorithms, the "lightgbm", "xgboost", and "e1071" packages were preferred, respectively [59–61]. The feature importances of all models were visualized using the "ggplot2" package [62].

Results

To estimate AQI, eight years daily air pollutant data (2016–2024) for Iğdır were used to compare different machine learning models. Table 4 presents descriptive statistics such as mean, standard deviation, minimum, and maximum values for all air quality criteria and meteorological features used in this study. For descriptive analysis and interpretation, daily measurements were aggregated into monthly averages, and the values shown in the table represent the mean monthly values across the eight-year dataset (2016–2024). This approach highlights seasonal variations and long-term trends. In the Table 4, *n* denotes the number of observations (sample sizes) used in the analysis within the scope of the statistical approach. For example, the average temperature in January was 1.67°C, with the lowest temperature being −11.8°C and the highest temperature being 7.7°C. In February, the average temperature increased slightly to 2.66°C. A significant rise in temperature was observed from March onwards, with the highest values recorded in July and August (27.28°C and 27.54°C). The temperature decreased from September onwards and was measured as 1.66°C in December (Table 4).

The precipitation parameter generally remained at low levels, with a maximum value of 29.1 mm measured in November. Relative humidity varied throughout the year, reaching its highest average of 75.44% in December. Wind speed generally remained low, with a maximum of 3.8 m s^{-1} measured in December. Wind direction generally varied between 200°-250° (Table 4).

The PM_{10} concentrations increased during the winter months and were determined to be 181.61 µg m^{-3} on average in January. The PM_{10} values decreased during the summer months but remained at high levels. SO_2, NO_2, and O_3 concentrations also showed seasonal variation, with SO_2 values increasing significantly during the winter months. The SO_2 average was measured as 13.05 µg m^{-3} in December (Table 4).

When evaluated regarding AQI, the highest values were seen in January (172.91) and November (168.16), indicating that air quality was negatively affected in winter months. During the summer months, AQI remained at lower levels. All these results reveal that air quality in the region is strongly affected by seasonal changes and deteriorates in winter months. This situation can be attributed to increased fossil fuel use and meteorological conditions in winter months. The study makes a significant contribution to determining measures that can be taken to improve air quality in the region (Table 4).

Fig 2 illustrates the correlation coefficients among environmental factors (temperature, precipitation, relative humidity, wind direction, wind speed, PM_{10}, SO_2, NO_2, and O_3) and their relationship with AQI. The coefficients range between −1 and +1, where the magnitude indicates the strength and the sign shows the direction of the relationship.

According to the correlation matrix, the highest correlation coefficient is between PM_{10} and AQI, with a coefficient of 0.81. In addition, a relatively high and positive correlation of 0.56 between NO_2 and AQI indicates that increases in NO_2 levels may be associated with increased AQI values. Similarly, the correlation coefficient 0.40 between SO_2 and AQI is also noteworthy. However, the correlation of −0.25 between temperature and AQI suggests that temperature increases may be associated with decreased AQI values. Lower correlation values are observed between other variables (e.g., precipitation, relative humidity, wind direction) and AQI, indicating that the effect of these variables on AQI may be less pronounced.

Hyperparameter optimization is a critical step in machine learning model development, as it directly influences predictive performance and generalization ability. Identifying the optimal parameter combinations ensures that the models are neither underfitted nor overfitted, thereby improving the reliability of AQI forecasts. Fig 3 presents the three-dimensional surface plots for the optimized model results on the train set, including the XGBoost (top), LightGBM (middle), and SVM (bottom) models. These plots illustrate the model performance under different hyperparameter combinations, where the x and y axes represent the hyperparameters, and the z-axis represents the corresponding performance metrics.

Table 4. Annual air pollutant and metrological factor dataset.

Parameters	Month	n	T	P	RH	WD	WS	PM_{10}	SO_2	NO_2	O_3	AQI
Mean	January	197	1.67	0.26	71.19	208.96	0.65	181.61	15.01	41.05	29	172.91
SD			4.58	1.1	11.21	149.83	0.45	109.95	11.3	13.97	11.95	74.23
Min			−11.8	0	41.2	1	0	14.12	2.41	7.69	4.7	18.59
Max			7.7	8.3	97.2	360	2.8	648.34	80.81	74.99	65.24	552.9
Mean	February	180	2.66	0.27	60.23	213.59	0.9	125.97	11.25	36.47	41.26	142.11
SD			5.1	1.91	13.89	147.04	0.63	72	6.62	15.85	15.22	56.95
Min			−10.9	0	29.7	1	0.2	17.11	2.98	6.69	12.63	22.55
Max			11.7	25.1	88	360	3.4	334.95	42.19	85.41	95	391.77
Mean	March	229	9.12	0.77	54.15	234.47	1.24	76.7	8.02	25.89	54.98	103.72
SD			3.62	2.8	13.93	136.51	0.7	51.67	3.91	11.67	16.77	63.26
Min			0.3	0	28.6	1	0.1	12.55	1.64	5.33	15.23	22.16
Max			19.4	27.5	92.2	360	3.4	335.37	28.71	58.3	115.04	325.47
Mean	April	206	14.72	0.85	50.09	204.06	1.19	65.86	5.67	21.71	58.73	95.63
SD			3.42	2.27	13.42	149.9	0.54	35.01	2.2	10.85	20.01	55.51
Min			5	0	26	1	0.2	10.23	0.65	6.64	16.65	25.33
Max			24.2	15.2	85.8	360	3.2	183.49	17.25	54.03	108.6	325.32
Mean	May	192	18.39	1.46	55.33	235.99	1.22	49.67	5.05	16.65	61.87	76.79
SD			3	2.84	13.12	141.29	0.49	25.89	2.26	6.39	22.46	52.17
Min			12.4	0	24.3	1	0.2	9.38	1.4	3.34	19.62	16.65
Max			26.7	19.3	84	360	3.6	142.35	12.14	36.55	115.96	304.69
Mean	June	161	24.13	0.77	47.33	244.76	1.25	64.35	5.13	15.82	73.32	99.47
SD			3.03	1.88	11.7	135.32	0.41	33.04	2.5	7.1	28.67	48.53
Min			17.3	0	20.8	1	0.4	12.4	1.6	4.77	16.29	27.01
Max			30.5	12.9	77.3	360	2.8	197.12	10.21	58.2	145.71	304.54
Mean	July	192	27.28	0.57	44.47	231.65	1.32	65.41	4.66	11.94	78.52	100.11
SD			2.55	1.69	8.79	147.7	0.52	35.81	2.06	3.67	28.38	49.72
Min			20.5	0	28.4	1	0.4	11.99	1.88	5.89	22.78	26.86
Max			33.1	14.8	78.2	360	3.1	189.75	11.45	23.36	137.95	325.14
Mean	August	211	27.54	0.19	42.89	214.02	1.06	85.85	4.96	15.35	70.09	128.83
SD			1.71	0.7	6.76	155.08	0.37	39.77	2.26	5.29	28.83	51.76
Min			21.1	0	28.5	1	0.2	25.07	1.76	5.76	15.43	34.87
Max			31.4	5.7	69.8	359	2.5	230.12	14.36	31.23	131.74	325.22
Mean	September	160	22.56	0.18	45.77	259.97	1.03	95	5.33	19.09	51.88	125.86
SD			3.2	0.98	8.67	129.88	0.57	52.45	3.31	7.01	20.53	54.63
Min			13.9	0	22.3	1	0.1	12.94	2.31	6.32	9.17	21.72
Max			29.8	10.3	81	360	2.8	234.37	15.8	41.11	92.53	325.26
Mean	October	137	14.46	0.66	63.35	229.42	0.59	94.87	5.57	27.78	30.57	118.27
SD			3.54	1.7	13.01	153.13	0.49	58.7	3.27	9.13	14.44	57.5
Min			5.1	0	36	1	0	12.22	2.1	6.73	7.43	16.46
Max			22.6	8.7	92.2	360	2.5	312.38	22.73	57.7	68.94	325.2
Mean	November	142	7.51	0.76	71.29	229.46	0.38	181.42	9.99	38.89	23.45	168.16
SD			3.14	3.05	11.72	158.82	0.36	89.11	6.29	11.56	13.35	52.11
Min			−1.7	0	38	1	0	14.06	2.4	7.62	4.38	22.05
Max			16.1	29.1	95.7	360	1.9	398.5	27.57	64.07	69.44	325.49
Mean	December	162	1.66	0.39	75.44	201.98	0.56	190.7	13.05	41.79	22.15	173.55
SD			4.36	1.58	12.47	158.47	0.57	105.89	7.39	16.83	13.03	59.69

(Continued)

Table 4. (Continued)

Parameters	Month	n	T	P	RH	WD	WS	PM_{10}	SO_2	NO_2	O_3	AQI
Min			−13.2	0	36.1	1	0	31.13	3.95	5.95	3.27	31.13
Max			10.6	15.9	98.5	360	3.8	545.13	46.58	125.03	66.78	510.02

https://doi.org/10.1371/journal.pone.0334252.t004

In the XGBoost section (top), the model performance shows relatively low variance across different hyperparameter settings, indicating stable results under specific parameter selections. The performance metric exhibits significant improvement and reaches peak values for certain hyperparameter combinations, highlighting the optimal conditions for the model (Fig 3). Additionally, error values are lower where the R^2 value is maximized, confirming the importance of model-specific hyperparameter tuning (Fig 3).

The LightGBM section (middle) includes three plots representing R^2, RMSE, and MAE metrics. The R^2 plot demonstrates how well the model explains variance in the dataset, with high values for specific parameter settings. The RMSE and MAE plots indicate the magnitude of prediction errors, where consistently low values across a wide range of hyperparameters suggest the model's ability to make accurate and reliable predictions (Fig 3).

The SVM section (bottom) also displays R^2, RMSE, and MAE values, providing critical insights for optimizing the hyperparameters of the SVM model. The variations in these metrics highlight the model's sensitivity to hyperparameter tuning, showing that proper selection of hyperparameters significantly impacts overall performance (Fig 3). The trends in these plots guide further optimization efforts to enhance the model's predictive accuracy.

This comprehensive visualization allows for a comparative evaluation of the models, facilitating a deeper understanding of their respective performances under different parameter settings (Fig 3).

Fig 4 presents the three-dimensional surface plots for the optimized model results on the test set, showcasing the XGBoost (top), LightGBM (middle), and SVM (bottom) models. These visualizations demonstrate how model performance varies with different hyperparameter settings, where the x- and y-axes correspond to the selected hyperparameters and the z-axis reflects the associated performance metric, thereby highlighting the sensitivity of each model to parameter tuning.

In the XGBoost section (top), the performance metrics—R^2, RMSE, and MAE—demonstrate a pattern similar to the training set results. The consistency between the train and test sets indicates that the model maintains stable predictive capabilities across different datasets. Specifically, the plots reveal that high R^2 values correspond to lower RMSE and MAE, confirming that the selected hyperparameter combinations contribute to accurate model predictions (Fig 4).

The LightGBM section (middle) also provides insights into the model's predictive success using R^2, RMSE, and MAE metrics. The R^2 plot highlights regions where the model effectively explains variance, particularly for certain hyperparameter settings. The RMSE and MAE plots indicate how prediction errors fluctuate under different conditions, with lower values signifying better performance (Fig 4). The model demonstrates robust generalization ability as it maintains low error rates across a broad range of hyperparameter settings.

The SVM section (bottom) displays performance variations based on different hyperparameter choices. The R^2 plot illustrates the model's explanatory power, while RMSE and MAE plots highlight fluctuations in prediction errors. Notably, the wave-like structures in the plots indicate that the model is highly sensitive to specific hyperparameter selections. This sensitivity emphasizes the importance of fine-tuning to achieve optimal performance. Additionally, the results provide valuable insights into the model's generalization capacity on test data, aiding in hyperparameter optimization (Fig 4).

By consolidating these models into a single Fig, Fig 3 facilitates a comparative evaluation, allowing for a clearer understanding of how each model performs on the test dataset under various hyperparameter settings.

Table 5 compares the goodness of fit criteria obtained with optimal hyperparameter values when LightGBM, XGBoost, and SVM algorithms are used in AQI estimation. LightGBM algorithm observed R^2 values as 0.922 and 0.889

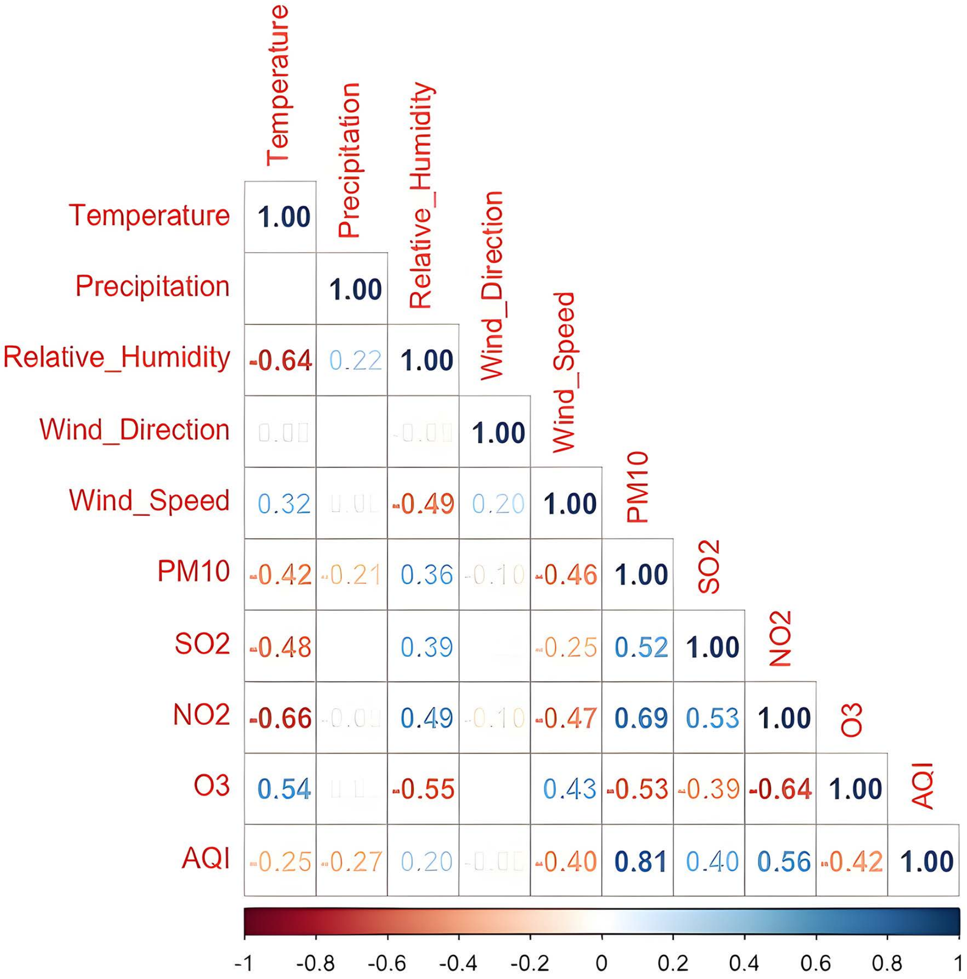

Fig 2. Correlation matrix for response and explanatory variables.

https://doi.org/10.1371/journal.pone.0334252.g002

in training and test datasets, respectively, indicating that the model can explain the variance in the dataset to a large extent. RMSE and MAE values for the training set were measured as 18.661 and 5.666, respectively, while for the test set, these values were measured as 20.764 and 6.777, indicating that the model has slightly higher error rates in the test set.

The XGBoost algorithm performs better than other models, with R^2 values indicating an almost perfect fit (0.999 in training, 0.994 in test). RMSE and MAE values are extremely low especially in the training set (RMSE 0.234, MAE 0.158), while these values are measured as 4.84 and 0.972 in the test set, suggesting that the model may have overfitted the training data. The SVM model showed an average fit with an R^2 value of 0.782, which remained constant in both training and test sets; RMSE and MAE values were determined as 28.824 and 12.233 in the training set and 31.136 and 13.546 in the test set. These results show that the SVM model performs less than the other two models on this particular data set. As a result, the XGBoost algorithm stands out as the best-performing model in AQI estimation with high R^2 values and low error metrics. The performances of LightGBM and SVM should be evaluated with more comprehensive analyses, especially in terms of generalization capabilities and error rates.

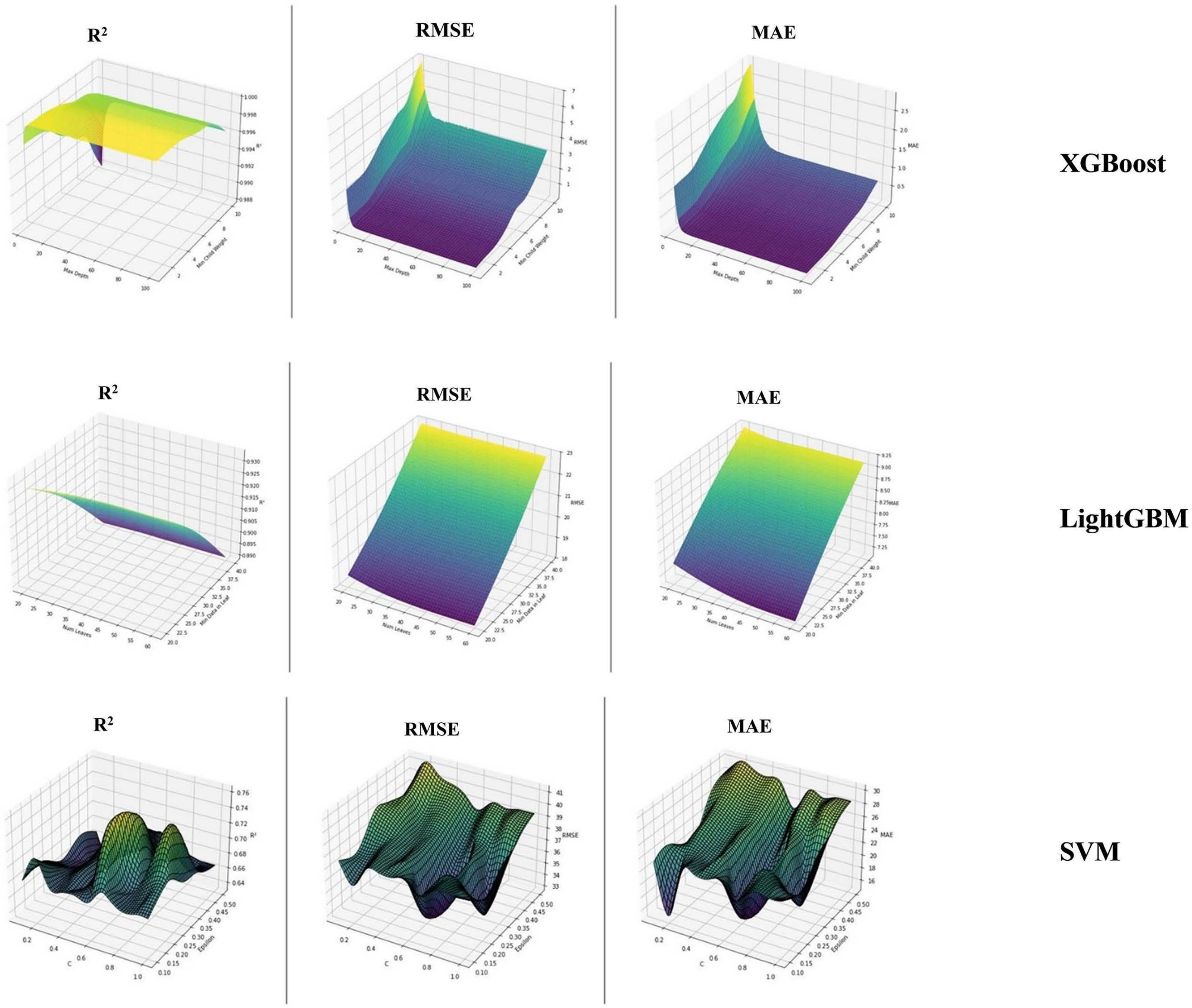

Fig 3. Optimized model results with 3D surface plot (Train set).

https://doi.org/10.1371/journal.pone.0334252.g003

Fig 5 presents the variable importance levels for the three models—XGBoost (Fig 5a), LightGBM (Fig 5b), and SVM (Fig 5c)—and their respective impacts on AQI estimates. These plots highlight the significance of different variables in predicting AQI, offering insights into how each model prioritizes features.

In the XGBoost section (Fig 5a), PM_{10} emerges as the most influential variable with an importance score of 0.974, followed by O_3, SO_2, and NO_2 with scores of 0.008, 0.007, and 0.004, respectively. Other features, such as Wind Direction, Relative Humidity, and Temperature, hold lower importance levels, indicating their relatively minor impact on AQI predictions.

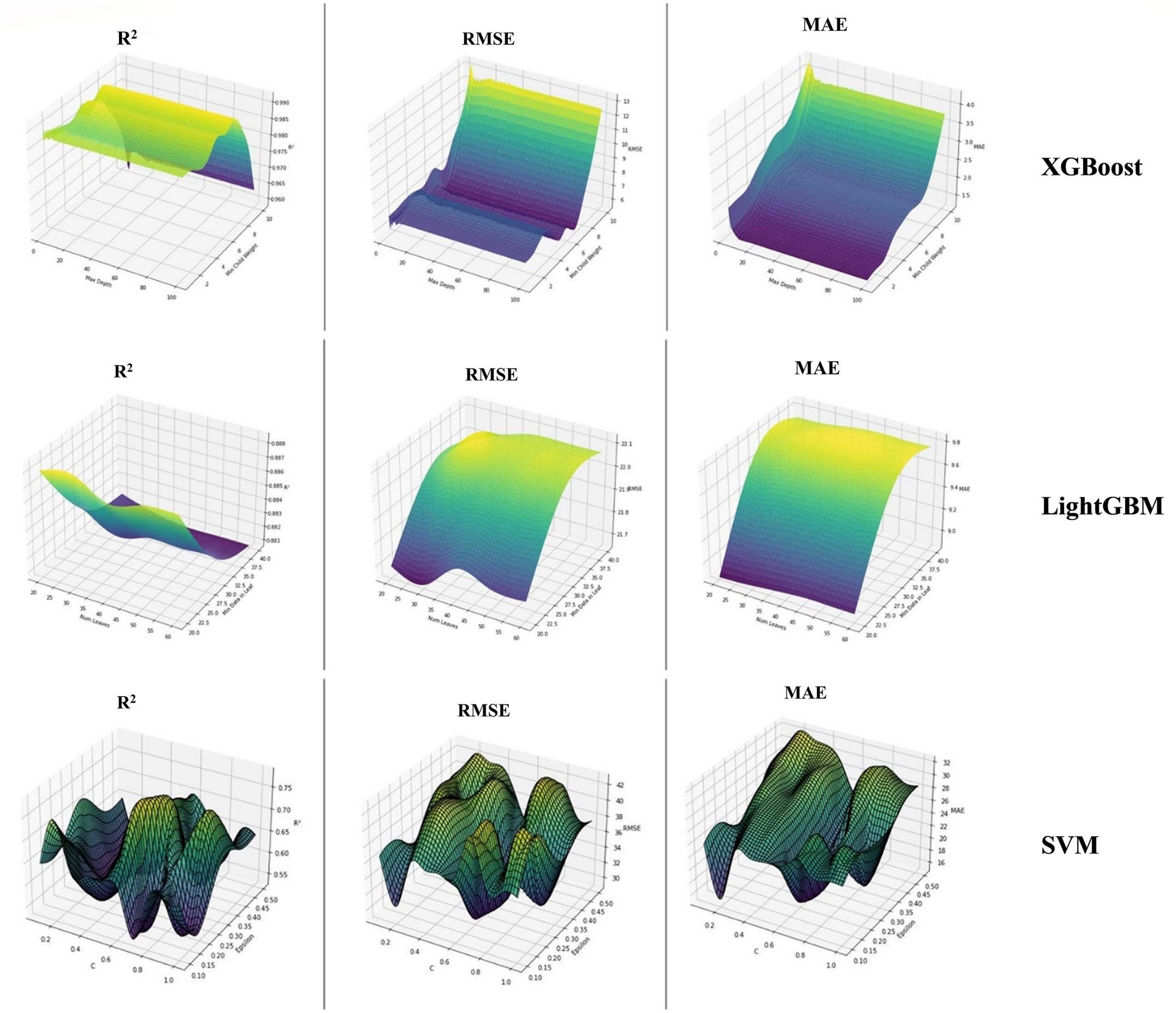

Fig 4. Optimized model results with 3D surface plot (Test set).

https://doi.org/10.1371/journal.pone.0334252.g004

The LightGBM section (Fig 5b) presents a variable importance distribution similar to that of XGBoost. PM_{10} remains the most critical predictor with a score of 0.955, while O_3, SO_2, and NO_2 follow with importance values of 0.011, 0.010, and 0.009, respectively. Wind Direction, Wind Speed, and Temperature are identified as less significant features in the AQI estimation process.

The SVM section (Fig 5c) exhibits a different variable importance distribution compared to the other models. In this case, PM_{10} maintains the highest importance with a score of 0.330, but the ranking of secondary variables shifts. Precipitation, Wind Speed, and Relative Humidity gain prominence, with scores of 0.103, 0.081, and 0.080, respectively.

Table 5. The goodness of fit criteria results in all algorithms for optimal hyperparameter values.

Hyperparameter values for each model

LightGBM		**XGBoost**		**SVM**	
Learning rate	0.04	eta	0.1	C	1
Num leaves	30	Max depth	10	sigma	0.1
Min data in leaf	20	Min child weight	1	kernel	radial

Goodness of fit criteria

	LightGBM		**XGBoost**		**SVM**	
Criterion	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
R^2	0.922	0.889	0.999	0.994	0.782	0.782
RMSE	18.661	20.764	0.234	4.84	28.824	31.136
MAE	5.666	6.777	0.158	0.972	12.233	13.546

https://doi.org/10.1371/journal.pone.0334252.t005

Conversely, Temperature, SO_2, and O_3 hold lower importance levels, suggesting their diminished role in AQI predictions within this model. This finding is consistent with Choudhary et al. [63], who reported that various machine learning algorithms such as RF, SVM, Bagged MARS, and BRNN provided reliable predictions of particulate matter and gaseous pollutants, supporting the overall effectiveness of ML-based approaches in air quality studies.

A comparative analysis of these models indicates that XGBoost and LightGBM yield highly similar variable importance rankings, identifying the same key predictors for AQI estimation. This alignment reinforces the reliability of these models in accurately capturing influential environmental factors. Meanwhile, SVM assigns relatively lower importance to PM_{10} and emphasizes different meteorological variables, reflecting variations in how models process and interpret input features.

By consolidating these variable importance analyses into a single Fig, Fig 5 facilitates a clearer comparison across models, offering a comprehensive understanding of how each algorithm prioritizes different features in AQI estimation.

In the present study, the XGBoost model exhibits impressive goodness of fit with R^2 values of 99.9% and 99.4% on the training and test sets, respectively, in various hyperparameter combinations. These high R^2 values indicate that the model explains the variance in the data set extraordinarily well. The RMSE and MAE values of the XGBoost model are much lower than the other models, indicating that the estimates are closer to the actual values. Therefore, the model is more reliable. While the LightGBM and SVM models also exhibit consistent results, the superior performance of the XGBoost model suggests that it should be preferred more, especially in terms of the accuracy of the estimates and the general reliability of the model. In this context, it is concluded that the XGBoost model offers higher reliability and accuracy when making AQI estimates.

Discussion

Monitoring and improving air quality is essential for both public authorities and individuals seeking to reduce environmental and health risks. Accurate AQI estimation supports these efforts, and machine learning algorithms provide powerful tools for this task. In this study, three algorithms—SVM, LightGBM, and XGBoost—were applied to AQI prediction using pollutant and meteorological data from Iğdır, Türkiye. All three models achieved satisfactory performance, but XGBoost consistently outperformed the others, yielding $R^2 = 0.999$, RMSE = 0.234, and MAE = 0.158. These results demonstrate that XGBoost is an effective and reliable model for AQI prediction, with reduced risk of overfitting compared to previous approaches.

These findings confirm that XGBoost has a superior ability to handle particularly complex datasets and understand the interactions between multidimensional features. Similar performance of XGBoost has also been reported in other studies [64,65]. For example, Van et al. [13] compared the performances of Decision Tree, Random Forest, and XGBoost

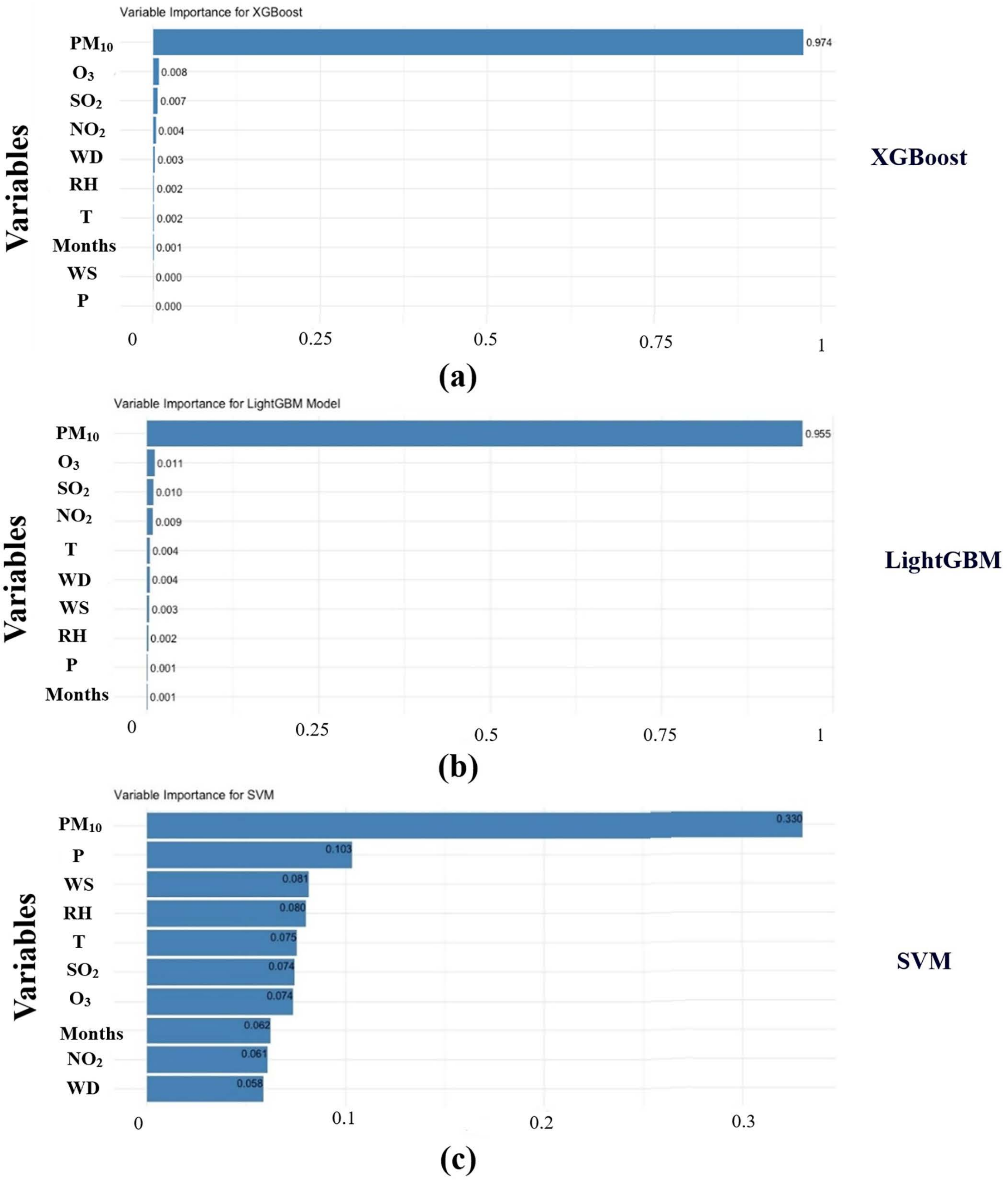

Fig 5. Variable importance for models. (a) XGBoost; (b) LightGBM; (c) SVM.

https://doi.org/10.1371/journal.pone.0334252.g005

algorithms. They stated that the XGBoost model gave the best results in accuracy ($R^2 = 0.9993$), error (RMSE = 2.5359), and MAE (1.2844) metrics on two different datasets.

Although the LightGBM model showed lower performance than XGBoost, it provided a good level of accuracy in AQI prediction. R^2 values for LightGBM were recorded as 0.922 in the training set and 0.889 in the test set. RMSE values were

18.661 and 20.764, and MAE values were 5.666 and 6.777, respectively. These findings show that LightGBM offers an important alternative for relatively less complex models. Ravindiran et al. [25], in particular, stated that CatBoost stood out in their similar studies, but LightGBM also produced reliable results in most cases.

The SVM model, on the other hand, exhibited lower performance compared to the other two models. The R^2 value was recorded as 0.782 in both training and test sets. The RMSE values were 28.824 and 31.136, and the MAE values were 12.233 and 13.546, respectively. These results indicate that SVM should be optimized for more complex datasets and multidimensional features. Liu et al. [33] emphasized that an SVR-based model was successful in some cases, but its generalizability may be limited.

The challenges associated with traditional air quality monitoring methods further emphasize the importance of using machine learning for AQI estimation. Traditional approaches often rely on fixed monitoring stations, which can provide limited spatial coverage and cannot effectively capture local pollution events [66,67]. In contrast, machine learning models play a pivotal role in integrating diverse datasets, including real-time sensor data and historical pollution records, to improve the accuracy of predictions. This comprehensive approach is particularly important in urban environments, where pollution sources can vary significantly in different areas [68,69]. Moreover, integrating meteorological parameters into predictive models is crucial because weather conditions such as temperature, humidity, and wind speed can significantly affect pollutant distribution and concentration levels [70,71]. The impact of meteorological parameters (temperature, precipitation, wind direction, wind speed, and humidity) on air quality is important in this study. The findings showed that meteorological factors are important inputs for AQI estimation, and air pollution varies depending on environmental conditions. Sigamani and Venkatesan [30] stated that meteorological factors play an important role in AQI estimation, affecting pollution concentrations by 60–74%.

This study's findings align with various machine learning-based AQI prediction studies in literature. Ravindiran et al. [24] highlighted that the CatBoost model performed best with $R^2 = 0.9998$ and $RMSE = 0.76$; however, in this study, the performance of XGBoost is close to CatBoost. Similarly, Liu et al. [33] found that an SVR-based model was superior in AQI prediction ($R^2 = 0.9766$), but the study did not cover newer algorithms such as XGBoost.

The study's findings also contribute to the ongoing discourse on overfitting in machine learning models. Overfitting occurs when a model learns from the noise in the training data rather than the underlying patterns, leading to poor generalization of unseen data. The ability of the XGBoost model to maintain consistent performance across a range of metric scores suggests that it effectively mitigates overfitting, a concern noted in previous research on machine learning applications in environmental science [72,73]. This feature is particularly valuable in the context of air quality prediction, where accurate prediction is essential for public health interventions and policymaking.

Moreover, the implications of improved AQI prediction extend beyond purely academic interest; there are real-world applications in public health and urban planning. Accurate AQI predictions can inform government responses to pollution events, enabling timely public health warnings and interventions. For example, during periods of high pollution, authorities can impose traffic restrictions or encourage public transportation to reduce exposure risks [74,75]. Additionally, individuals can use AQI predictions to make informed decisions about outdoor activities, reducing the health risks associated with poor air quality [76–77]. In this regard, Kumar et al. [78] further noted that conventional AQI measures may underestimate actual health risks, highlighting the importance of developing reliable forecasting systems that can better inform public protection strategies.

The study's results are significant, particularly in the context of recent global events such as the COVID-19 pandemic. During the pandemic, several studies demonstrated that poor air quality aggravated respiratory conditions and increased vulnerability to severe outcomes from SARS-CoV-2 infection [79,80]. In this regard, the ability to accurately predict AQI is especially valuable, as it can support early interventions and public health strategies aimed at reducing exposure to harmful pollutants during health crises. This makes the study's findings highly relevant for protecting public health, particularly in densely populated urban areas where pollution levels are often elevated.

Most of the studies summarized in Table 2 focus on short-term datasets, a limited number of pollutant variables, or exclusively metropolitan regions. Moreover, many of these studies do not incorporate the long-term combined evaluation of meteorological and pollutant parameters. In this context, our study contributes to literature by using a long-term dataset that integrates both pollutant and meteorological variables. Conducting such a long-term analysis in a geopolitically sensitive and climatically unique region like Iğdır further reinforces the originality and significance of our findings.

In conclusion, the application of machine learning algorithms, particularly XGBoost, to predict AQI in Iğdır, Türkiye, demonstrates the potential of these technologies to improve air quality management. The study's findings not only highlight the effectiveness of machine learning in this area but also highlight the importance of integrating diverse datasets and addressing overfitting to increase model reliability. As urbanization continues to increase and air pollution remains a pressing global issue, developing robust predictive models will be important to create healthier environments and support public well-being.

Conclusion

Air pollution levels in Iğdır have increased significantly, particularly during the winter months. The average AQI values in January and November were 172.9 and 168.2, respectively, which correspond mostly to the "Unhealthy" category according to the TNAQI classification. PM_{10} was determined to be the primary pollutant in pollution.

Analysis and model comparison results indicate that the XGBoost model achieved significantly superior performance in AQI predictions compared to the LightGBM and SVM models. While the R^2 values of the XGBoost model were exceptionally high in both the training (99.9%) and test (99.4%) sets, the error metrics, such as RMSE and MAE, were also notable with low values such as 0.234 and 0.158, respectively. These results show that the model explains the variance in the dataset perfectly well, and the predictions are highly accurate.

On the other hand, the LightGBM model also exhibited robust results, with R^2 values measured as 92.2% and 88.9% in the training and test sets, respectively. This model also provided acceptable performance with low RMSE and MAE values. On the other hand, the SVM model exhibited lower reliability than these two models, with R^2 values remaining at 78.2% in both sets and relatively high RMSE and MAE values, indicating that the model is not as effective as the others.

Considering these findings, the XGBoost model offers a more reliable and effective alternative for AQI estimations than other models. These results underline XGBoost's high adaptability and accuracy capacity, especially in complex data structures and when evaluating the interaction of various environmental parameters. The proposed approach can serve as a valuable guide for future modeling studies and constitute a basis for applications on larger data sets.

Acknowledgments

The author is grateful to the 'Republic of Türkiye Ministry of Environment, Urbanization and Climate Change' and for providing the air quality and meteorological data.

Author contributions

Conceptualization: Sevtap Tırınk.

Data curation: Sevtap Tırınk.

Formal analysis: Sevtap Tırınk.

Investigation: Sevtap Tırınk.

Methodology: Sevtap Tırınk.

Resources: Sevtap Tırınk.

Software: Sevtap Tırınk.

Supervision: Sevtap Tırınk.

Validation: Sevtap Tırınk.

Visualization: Sevtap Tırınk.

Writing – original draft: Sevtap Tırınk.

Writing – review & editing: Sevtap Tırınk.

References

1. Babatola SS. Global burden of diseases attributable to air pollution. J Public Health Afr. 2018;9(3):813. https://doi.org/10.4081/jphia.2018.813 PMID: 30687484
2. Kumar A, Patil RS, Dikshit AK, Kumar R. Comparison of predicted vehicular pollution concentration with air quality standards for different time periods. Clean Techn Environ Policy. 2016;18(7):2293–303. https://doi.org/10.1007/s10098-016-1147-6
3. Dadkhah-Aghdash H, Rasouli M, Rasouli K, Salimi A. Detection of urban trees sensitivity to air pollution using physiological and biochemical leaf traits in Tehran, Iran. Sci Rep. 2022;12(1):15398. https://doi.org/10.1038/s41598-022-19865-3 PMID: 36100647
4. Zhou C, Li S, Wang S. Examining the impacts of urban form on air pollution in developing countries: a case study of China's megacities. Int J Environ Res Public Health. 2018;15(8):1565. https://doi.org/10.3390/ijerph15081565 PMID: 30042324
5. Patton AP, Perkins J, Zamore W, Levy JI, Brugge D, Durant JL. Spatial and temporal differences in traffic-related air pollution in three urban neighborhoods near an interstate highway. Atmos Environ (1994). 2014;99:309–21. https://doi.org/10.1016/j.atmosenv.2014.09.072 PMID: 25364295
6. Argun YA, Tırınk S, Bayram T. Effect of urban factors on air pollution of Igdir. Black Sea J Eng Sci. 2019;2(4):123–30. https://doi.org/10.34248/bsengineering.561588
7. Tırınk S, Öztürk B. Evaluation of PM10 concentration by using Mars and XGBOOST algorithms in Iğdır Province of Türkiye. Int J Environ Sci Technol. 2023;20:5349–58.
8. Tui Y, Qiu J, Wang J, Fang C. Analysis of spatio-temporal variation characteristics of main air pollutants in Shijiazhuang city. Sustainability. 2021;13(2):941. https://doi.org/10.3390/su13020941
9. Erawan M, Karuniasa M. Spatial dynamics of air pollution in Tangerang City, Jabodetabek metropolitan area. In: Proceedings of the 13th International Interdisciplinary Studies Seminar, IISS 2019. 30–1.
10. Kumar RP, Prakash A, Singh R, Kumar P. Machine learning-based prediction of hazards fine PM2.5 concentrations: a case study of Delhi, India. Discov Geosci. 2024;2(1). https://doi.org/10.1007/s44288-024-00043-z
11. Cho H-S, Choi M. Effects of compact urban development on air pollution: empirical evidence from Korea. Sustainability. 2014;6(9):5968–82. https://doi.org/10.3390/su6095968
12. Abas A, Aiyub K, Awang A. Biomonitoring potentially toxic elements (PTES) using lichen transplant usnea misaminensis: a case study from Malaysia. Sustainability. 2022;14(12):7254. https://doi.org/10.3390/su14127254
13. Van NH, Van Thanh P, Tran DN, Tran DT. A new model of air quality prediction using lightweight machine learning. Int J Environ Sci Technol. 2023;20:2983–94.
14. EPA. National air quality and emissions trends report 1997. 454/R98-016. EPA; 1997.
15. EPA. Air quality index reporting; final rule. Fed Reg. CFR; 1999;Part III.
16. TNAQI. Republic of Türkiye ministry of environment, urbanisation and climate change. the Turkish national air quality index (TNAQI). 2025. https://dathm.csb.gov.tr/hava-kalitesi-indeksi-i-89066
17. Yang Q, Yuan Q, Li T, Shen H, Zhang L. The Relationships between PM2.5 and meteorological factors in China: seasonal and regional variations. Int J Environ Res Public Health. 2017;14(12):1510. https://doi.org/10.3390/ijerph14121510 PMID: 29206181
18. Sekula P, Ustrnul Z, Bokwa A, Bochenek B, Zimnoch M. Random forests assessment of the role of atmospheric circulation in PM10 in an urban area with complex topography. Sustainability. 2022;14(6):3388. https://doi.org/10.3390/su14063388
19. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. 2017;17(1):907. https://doi.org/10.1186/s12889-017-4914-3 PMID: 29179711
20. Liu Y, Wang P, Li Y, Wen L, Deng X. Air quality prediction models based on meteorological factors and real-time data of industrial waste gas. Sci Rep. 2022;12(1):9253. https://doi.org/10.1038/s41598-022-13579-2 PMID: 35661145
21. Zhang B, Duan M, Sun Y, Lyu Y, Hou Y, Tan T. Air quality index prediction in six major chinese urban agglomerations: a comparative study of single machine learning model, ensemble model, and hybrid model. Atmosphere. 2023;14(10):1478. https://doi.org/10.3390/atmos14101478
22. Ansari A, Quaff AR. Advanced machine learning techniques for precise hourly air quality index (AQI) prediction in Azamgarh, India. Int. J. Environ. Res. 2025;19, 1–31.
23. Aram SA, Nketiah EA, Saalidong BM, Wang H, Afitiri AR, Akoto AB, et al. Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. Int J Environ Sci Technol. 2024;21:1345–60.

24. Ravindiran G, Hayder G, Kanagarathinam K, Alagumalai A, Sonne C. Air quality prediction by machine learning models: a predictive study on the indian coastal city of Visakhapatnam. Chemosphere. 2023;338:139518. https://doi.org/10.1016/j.chemosphere.2023.139518 PMID: 37454985

25. Gupta NS, Mohta Y, Heda K, Armaan R, Valarmathi B, Arulkumaran G. Prediction of air quality index using machine learning techniques: a comparative analysis. J Environ Public Health. 2023;2023:1–26. https://doi.org/10.1155/2023/4916267

26. Maltare NN, Vahora S. Air quality index prediction using machine learning for Ahmedabad city. Digit Chem Eng. 2023;7:1–9.

27. Zhang H, Ge L, Zhang G, Fan J, Li D, Xu C. A two-stage intrusion detection method based on light gradient boosting machine and autoencoder. Math Biosci Eng. 2023;20(4):6966–92. https://doi.org/10.3934/mbe.2023301 PMID: 37161137

28. Sarkar N, Gupta R, Keserwani PK, Govil MC. Air quality index prediction using an effective hybrid deep learning model. Environ Pollut. 2022;315:120404. https://doi.org/10.1016/j.envpol.2022.120404 PMID: 36240962

29. Pant A, Sharma S, Bansal M, Narang M. Comparative analysis of supervised machine learning techniques for AQI prediction. In: 2022 International conference on advanced computing technologies and applications (ICACTA), 2022. 1–4. https://doi.org/10.1109/icacta54488.2022.9753636

30. Sigamani S, Venkatesan R. Air quality index prediction with influence of meteorological parameters using machine learning model for IoT application. Arab J Geosci. 2022;15:1–12.

31. Janarthanan R, Partheeban P, Somasundaram K, Navin Elamparithi P. A deep learning approach for prediction of air quality index in a metropolitan city. Sustain Cities Soc. 2021;67:102720. https://doi.org/10.1016/j.scs.2021.102720

32. Castelli M, Clemente FM, Popovič A, Silva S, Vanneschi L. A machine learning approach to predict air quality in California. Complexity. 2020;2020:1–23. https://doi.org/10.1155/2020/8049504

33. Liu H, Li Q, Yu D, Gu Y. Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci. 2019;9(19):1–9.

34. Nebenzal A, Fishbain B. Long-term forecasting of nitrogen dioxide ambient levels in metropolitan areas using the discrete-time Markov model. Environ Model Softw. 2018;107:175–85. https://doi.org/10.1016/j.envsoft.2018.06.001

35. Liu D, Lee S, Huang Y, Chiu C. Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. Expert Syst. 2019;37(3). https://doi.org/10.1111/exsy.12511

36. Liang Y-C, Maimury Y, Chen AH-L, Juarez JRC. Machine learning-based prediction of air quality. Appl Sci. 2020;10(24):9151. https://doi.org/10.3390/app10249151

37. Shen J, Valagolam D, McCalla S. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. PeerJ. 2020;8:e9961. https://doi.org/10.7717/peerj.9961 PMID: 32983651

38. Dairi A, Harrou F, Khadraoui S, Sun Y. Integrated multiple directed attention-based deep learning for improved air pollution forecasting. IEEE Trans Instrum Meas. 2021;70:1–15. https://doi.org/10.1109/tim.2021.3091511

39. Pappa A, Kioutsioukis I. Forecasting particulate pollution in an urban area: from copernicus to Sub-Km scale. Atmosphere. 2021;12(7):881. https://doi.org/10.3390/atmos12070881

40. Ke G, Meng Q, Finley T. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st conference on neural information processing systems (NeurIPS 2017), Long Beach, CA, USA, 2017.

41. Li F, Zhang L, Chen B. A light gradient boosting machine for remaining useful life estimation of aircraft engines. In: Proceedings of the International conference on intelligent transportation, Maui, HI, USA, 2018.

42. Hajihosseinlou M, Maghsoudi A, Ghezelbash R. A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. Nat Resour Res. 2023;32:2417–38.

43. Cai J, Li X, Tan Z, Peng S. An assembly-level neutronic calculation method based on LightGBM algorithm. Annal Nuclear Energy. 2021;150:107871. https://doi.org/10.1016/j.anucene.2020.107871

44. Zhang H, Ge L, Wang Z. A high-performance intrusion detection system using LightGBM based on oversampling and undersampling. In: International conference on intelligent computing, 2022.

45. Xiao X, Shao YT, Luo ZT, Qiu WR. m5C-HPromoter: an ensemble deep learning predictor for identifying 5-methylcytosine sites in human promoters. Curr Bioinform. 2022;5:452–61.

46. Cui B, Ye Z, Zhao H, Renqing Z, Meng L, Yang Y. Used car price prediction based on the iterative framework of XGBoost+LightGBM. Electronics. 2022;11(18):2932. https://doi.org/10.3390/electronics11182932

47. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, San Francisco, CA, USA, 2016. 785–94.

48. Szczepanek R. Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. Hydrology. 2022;9(12):226. https://doi.org/10.3390/hydrology9120226

49. Faraz A, Tırınk C, Önder H, Şen U, Ishaq HM, Tauqir NA, et al. Usage of the XGBoost and MARS algorithms for predicting body weight in Kajli sheep breed. Trop Anim Health Prod. 2023;55(4):276. https://doi.org/10.1007/s11250-023-03700-6 PMID: 37500805

50. Yang Y, Wu Y, Wang P, Jiali X. Stock price prediction based on XGBoost and LightGBM. E3S Web Conf. 2021;275:01040. https://doi.org/10.1051/e3sconf/202127501040

51. Nguyen QT, Fouchereau R, Frénod E, Gerard C, Sincholle V. Comparison of forecast models of production of dairy cows combining animal and diet parameters. Comput Electron Agric. 2020;170:1–10.

52. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199–222.

53. Kavaklıoğlu K. Modeling and prediction of Turkey's electricity consumption using support vector regression. Appl Energy. 2011;88:368–75.

54. Patel AK, Chatterjee S, Gorai AK. Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades. Earth Sci Inform. 2018;12(2):197–210. https://doi.org/10.1007/s12145-018-0370-6

55. R Core Team. R: a language and environment for statistical computing; R foundation for statistical computing: Vienna, Austria, 2022. https://www.r-project.org/

56. Wei T, Simko V. Visualization of a correlation matrix. R Package; 2021.

57. Revelle W. Psych: procedures for personality and psychological research. Evanston, IL, USA: Northwestern Scholars; 2022.

58. Kuhn M. Caret: classification and regression training. Astrophysics Data System; 2022.

59. Chen T, He T, Benesty M. xgboost: extreme gradient boosting. 2023.

60. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. 2024.

61. Shi Y, Ke G, Soukhavong D. Lightgbm: light gradient boosting machine. 2023.

62. Wickham H. ggplot2: elegant graphics for data analysis. New York, NY, USA: Springer; 2016.

63. Choudhary A, Kumar P, Pradhan C, Sahu SK, Chaudhary SK, Joshi PK, et al. Evaluating air quality and criteria pollutants prediction disparities by data mining along a stretch of urban-rural agglomeration includes coal-mine belts and thermal power plants. Front Environ Sci. 2023;11. https://doi.org/10.3389/fenvs.2023.1132159

64. Srivastava S, Kumar A, Bauddh K, Gautam AS, Kumar S. 21-day lockdown in india dramatically reduced air pollution indices in Lucknow and New Delhi, India. Bull Environ Contam Toxicol. 2020;105(1):9–17. https://doi.org/10.1007/s00128-020-02895-w PMID: 32495123

65. Benchrif A, Wheida A, Tahri M, Shubbar RM, Biswas B. Air quality during three covid-19 lockdown phases: AQI, PM2.5 and NO2 assessment in cities with more than 1 million inhabitants. Sustain Cities Soc. 2021;74:103170. https://doi.org/10.1016/j.scs.2021.103170 PMID: 34290956

66. Zhang Z, Xue T, Jin X. Effects of meteorological conditions and air pollution on COVID-19 transmission: evidence from 219 Chinese cities. Sci Total Environ. 2020;741:140244. https://doi.org/10.1016/j.scitotenv.2020.140244 PMID: 32592975

67. Sarroeira R, Henriques J, Sousa AM, Ferreira da Silva C, Nunes N, Moro S, et al. Monitoring sensors for urban air quality: the case of the municipality of Lisbon. Sensors (Basel). 2023;23(18):7702. https://doi.org/10.3390/s23187702 PMID: 37765759

68. Dong D, Xu X, Yu H, Zhao Y. The impact of air pollution on domestic tourism in China: a spatial econometric analysis. Sustainability. 2019;11(15):4148. https://doi.org/10.3390/su11154148

69. Ethan CJ, Mokoena KK, Yu Y. Air pollution status in 10 mega-cities in China during the initial phase of the COVID-19 outbreak. Int J Environ Res Public Health. 2021;18(6):3172. https://doi.org/10.3390/ijerph18063172 PMID: 33808577

70. Kim B. Do air quality alerts affect household migration?. Southern Economic Journal. 2018;85(3):766–95. https://doi.org/10.1002/soej.12310

71. Wu H, Zhang Y, Yu Q, Ma W. Application of an integrated Weather Research and Forecasting (WRF)/CALPUFF modeling tool for source apportionment of atmospheric pollutants for air quality management: a case study in the urban area of Benxi, China. J Air Waste Manag Assoc. 2018;68(4):347–68. https://doi.org/10.1080/10962247.2017.1391009 PMID: 29020513

72. Balakrishnan K, Dey S, Gupta T, Dhaliwal RS, Brauer M, Cohen AJ, et. al. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the global burden of disease study 2017. Lancet Planet Health. 2019;3(1):e26–39. https://doi.org/10.1016/S2542-5196(18)30261-4 PMID: 30528905

73. Graça D, Reis J, Gama C, Monteiro A, Rodrigues V, Rebelo M, et al. Sensors network as an added value for the characterization of spatial and temporal air quality patterns at the urban scale. Sensors (Basel). 2023;23(4):1859. https://doi.org/10.3390/s23041859 PMID: 36850456

74. Rani N, Azid A, Khalit S, Juahir H, Samsudin M. Air pollution index trend analysis in Malaysia, 2010–15. Pol J Environ Stud. 2018;27:801–7.

75. Ming W, Zhou Z, Ai H, Bi H, Zhong Y. COVID-19 and air quality: evidence from China. Emerg Markets Fin Trade. 2020;56(10):2422–42. https://doi.org/10.1080/1540496x.2020.1790353

76. Gao H, Yang W, Yang Y, Yuan G. Analysis of the air quality and the effect of governance policies in China's pearl river delta, 2015–2018. Atmosphere. 2019;10(7):412. https://doi.org/10.3390/atmos10070412

77. Han C, Xu R, Zhang Y, Yu W, Zhang Z, Morawska L, et al. Air pollution control efficacy and health impacts: a global observational study from 2000 to 2016. Environ Pollut. 2021;287:1–9.

78. Kumar P, Choudhary A, Joshi PK, Kumar RP, Bhatla R. Machine learning models for estimating criteria pollutants and health risk-based air quality indices over eastern coast coal mine complex belts. Front Environ Sci. 2025;13. https://doi.org/10.3389/fenvs.2025.1589991

79. Ignac-Nowicka J. Towards smart city: influence of air pollution on the local community of the Zabrze city in surveys and field research. Multidiscip Asp Prod Eng. 2018;1:845–50.

80. Gurajala S, Dhaniyala S, Matthews JN. Understanding public response to air quality using tweet analysis. Social Media Soc. 2019;5(3). https://doi.org/10.1177/2056305119867656