ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

Air Quality Prediction in Shaanxi Province: A GBDT Model Study from the Digital-Intelligent Governance Perspective

Zhen Ma[1], Zihao Wang[2]

[1] School of Data Science, Xi'an Eurasia University, Xi'an 710065, China

[2] School of Data Science, Xi'an Eurasia University, Xi'an 710065, China

Abstract

This study constructs an air quality machine learning prediction model based on air quality monitoring and meteorological index data from Shaanxi Province, drawing on the concept of environmental digital intelligence governance. After data cleaning, feature extraction, and standardization processing, the performance of the decision tree, random forest, and GBDT algorithm is compared. The results show that the GBDT model has the best prediction effect, with an R^2 of 0.9832, which can effectively capture air quality changes, and PM10, PM2.5, and SO_2 are the main air quality influencing factors. The research results can provide data support for regional environmental protection and precise pollution control, and also provide reference cases for the practical application of digital intelligence technology in environmental governance.

Keywords

GBDT; Air Quality Forecasting; Environmental Digital Intelligence Governance.

1. Introduction

With the deepening of environmental digital intelligence governance, accurate air quality prediction has become the key to regional air pollution prevention and control. Shaanxi Province is an ideal area for verifying digital intelligence governance models due to its terrain in the Guanzhong Basin, which is not conducive to the diffusion of pollutants, and as an important industrial energy town in the west. In August 2025, the "Implementation Opinions of the Shaanxi Provincial People's Government on Deepening Air Pollution Control and Promoting the Achievement of the 14th Five-Year Plan Air Quality Goals" was released, further emphasizing the policy requirements for deepening control and achieving air quality goals [1].

Specifically, air quality digital intelligence prediction refers to the use of big data, machine learning, and other technologies to integrate and analyze multi-source environmental monitoring data to achieve accurate simulation and early research and judgment of pollution processes. Relevant studies at home and abroad show different technical prediction models. Foreign research focuses more on the innovation of basic models and the integration of interdisciplinary methods. For example, Luo Z et al. (2022) improved the prediction accuracy at the monitoring point scale by constructing a quadratic prediction ELM model [2]. Duan J et al. (2023) developed a hybrid framework combining ARIMA and CNN-LSTM, and introduced an intelligent optimization algorithm for parameter search [3]. Hua V et al. (2024) systematically compared the effects of multiple data filling methods on prediction performance [4]. Domestic research focuses more on the optimization and improvement of specific models and local practices, mainly focusing on deep learning model architecture. For example, Gao Morenhai et al. (2024) proposed the CNN-LSTM-Multi-Head Mechanisms model [5], and Zhou et al. (2025) constructed an SD-LSTM-ELM hybrid model based on quadratic decomposition and error correction [6]. Comprehensive analysis shows that although the existing

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

research has been deepening in terms of model complexity and accuracy, there are two limitations. First, the mainstream research is highly focused on deep learning models such as CNN, and the exploration of ensemble learning algorithms such as GBDT with strong interpretation, good at handling mixed features, and efficient computational efficiency is insufficient. Second, most research concentrates on model optimization itself, overlooking how predictive models can contribute to environmental governance within the broader context of environmental digital-intelligence governance.

Therefore, this study focuses on air quality prediction in Shaanxi Province by developing an integrated model that combines meteorological and pollutant monitoring data. The prediction performance of decision tree, random forest, and GBDT algorithms is compared, confirming the advantage of GBDT in capturing nonlinear air quality variations. Key influencing factors are identified through feature importance analysis. The results aim to support differentiated and refined environmental digital governance with actionable insights, while also offering an interpretable case study for applying machine learning in regional environmental management.

2. Model Basis and Evaluation Indicators

2.1 Model Basics

The decision tree is a typical supervised learning method, and its structure presents a tree-like topology form, relying on the recursive segmentation strategy of feature space to realize the progressive classification and regression prediction of samples. The theoretical basis of the algorithm is derived from Shannon information theory, and its node division mechanism is based on the information entropy minimization criterion. The classification boundary is constructed through recursive feature selection.

As a classical algorithm under the ensemble learning framework, the core mechanism of random forest relies on the self-service resampling technology in the Bagging algorithm to construct several differentiated decision tree-based classifiers by randomly selecting subsets in the feature space. When dealing with classification problems, the model follows the principle of collective decision-making, synthesizes the output results of each base classifier, and uses the majority voting mechanism to generate the final prediction.

The gradient lifting decision tree (GBDT) is an iterative optimization model that fits the negative gradient residuals of the previous round of models

$$(r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}) \tag{1}$$

Gradually build new trees. Taking the square loss function $(L(y, f(x)) = (y - f(x))^2)$ as an example, the model updates the formula to optimize forecasting capabilities with high accuracy and interpretability.

$$(f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{mj} I(x \in R_{mj})) \tag{2}$$

2.2 Evaluation Indicators

In the field of statistical modeling, prediction accuracy evaluation indicators usually include mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2).

The mean square error (MSE) is the square mean of the error between the predicted value and the true value, which can reflect the average error degree between the predicted value and the true value of the model, and the smaller the index value, the better. Root mean square error (RMSE) is a classic

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

error measure, and in the regression model evaluation system, RMSE is suitable for application scenarios that need to focus on monitoring significant errors, and the smaller the better. Mean absolute error (MAE) refers to the degree of discreteness of the predicted result by calculating the mean absolute deviation between the predicted value and the observed value. When the MAE measurement value approaches zero, the agreement between the prediction results and the actual observation value is significantly improved. The coefficient of determination (R^2) is an important goodness-of-fit index, and its numerical domain is limited to the [0,1] interval, and the fitting effect is evaluated by explaining the degree of data variation by the quantitative model. When the value approaches 1, it reflects that the model has excellent interpretive ability.

3. Research Implementation and Analysis

3.1 Data Collection and Preprocessing

This study uses public air quality monitoring data from Shaanxi Province from June 2024 to June 2025, with a total of 3,540 records. The data cover the concentrations of PM2.5, PM10, SO_2, NO_2, O_3 and CO major pollutants and the air quality index AQI, which provides sufficient support for the analysis of pollution trends, as shown in Table 1.

Table 1. Variable description

Variable	Variable Name	Illustration	Data Type	Units
Independent variable	PM2.5	The concentration of particulate matter in the atmosphere with a diameter of less than or equal to 2.5 microns, one of the key indicators used to measure air quality,	Numerical	ug
	PM10	The concentration of particulate matter in the atmosphere with a diameter of less than or equal to 10 microns reflects the amount of respirable particulate matter in the air	Numerical	ug
	SO_2	Sulfur dioxide concentration, mainly derived from the combustion of sulfur-containing fuels, is one of the air pollutants	Numerical	ug
	NO_2	Nitrogen dioxide concentration, mainly produced by motor vehicle exhaust and industrial exhaust emissions, has an impact on air quality and human health	Numerical	ug
	O_3	Ozone concentration, which is harmful to the human body when high concentrations near the ground, is the main component of photochemical smoke	Numerical	ug
	CO	Carbon monoxide concentration, mainly produced by incomplete combustion, is a toxic gas	Numerical	ug
Dependent variable	AQI	Air Quality Index	Numerical	0-100

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

Data cleaning was performed to ensure data quality by handling missing values, outliers, and duplicate entries. Specifically, methods included multiple imputation, box-plot detection for outlier replacement with the median, and the drop_duplicates() function. This process yielded a final dataset of 3,540 samples for analysis.

3.2 Data Exploratory Analysis

Before constructing the model, the dependent variable AQI, and variable correlations were analyzed. As shown in Fig.1, the AQI is mainly concentrated in the heavy pollution range of 250-300, and the span is 50 to 400 or more, and the air quality fluctuates greatly, which is easily affected by seasonal, industrial, and meteorological factors. The distribution is in a right-biased pattern, which may be related to industrial accidents or sandstorms and other pollution emergencies.

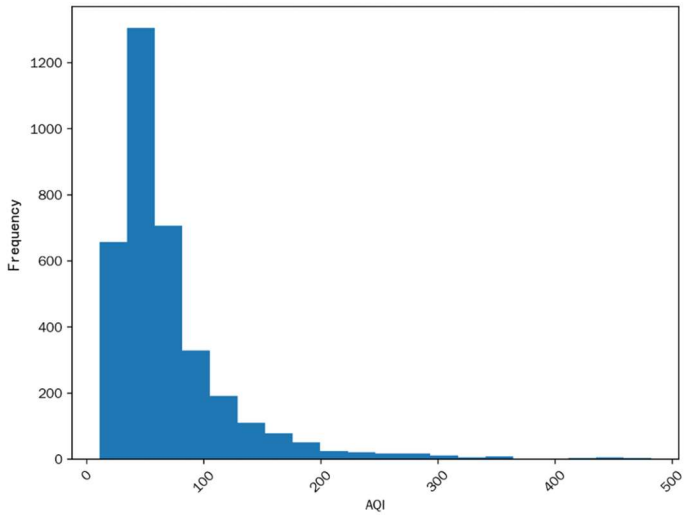

Fig.1 Histogram of AQI detection

Correlation analysis (seen in Fig.2) showed that the dependent variable AQI was highly correlated with PM2.5 and PM10 (both correlation coefficients were 0.91), indicating that the two were greatly influenced by each other, while O_3 was mostly negatively correlated with other variables, suggesting that the formation mechanism of O_3 and other pollutants may have opposite processes

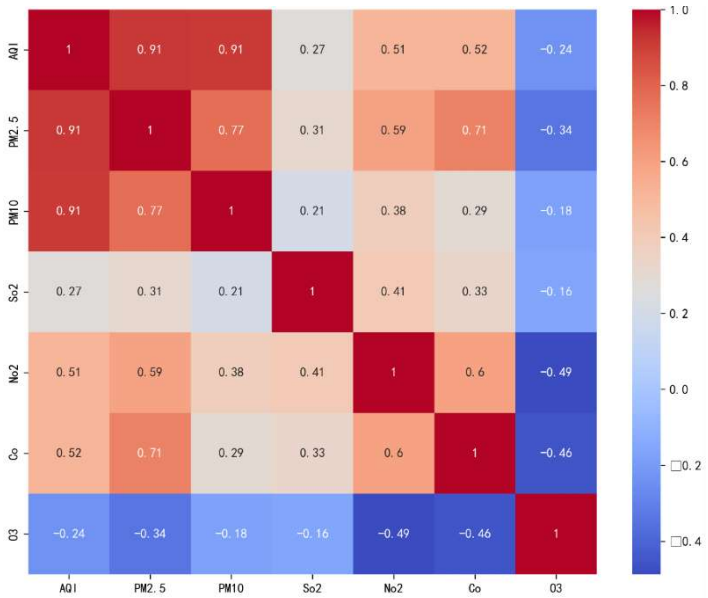

Fig. 2 Variable correlation heat map

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

3.3 Feature Engineering

The first step is to standardize the data. In order to eliminate the dimensional differences between different features and improve the training efficiency and accuracy of the model, the data is standardized and the data is mapped to the [0, 1] interval by using the Min-Max standardization method, and the data after data normalization is shown in Table 2.

Table 2. Min - Max standardized data

	AQI	PM2.5	PM10	SO_2	NO_2	CO	O_3
1	0.556263	0.671733	0.1968	0.232558	0.797753	0.570370	0.093960
2	0.498938	0.589666	0.1672	0.186047	0.730337	0.514815	0.167785
3	0.558386	0.674772	0.1928	0.162791	0.808989	0.585185	0.147651
4	0.636943	0.787234	0.2264	0.162791	0.943820	0.655556	0.093960
5	0.707006	0.887538	0.2608	0.162791	1.000000	0.703704	0.147651
...	...	...	...	...	...	...	...

Subsequently, the data is split. Using the train_test_split() function of the scikit-learn library in Python, the dataset is divided into a training set and a validation set according to the ratio of 70% and 30%.

3.4 Model Building and Predictive Analysis

3.4.1 Model Construction and Comparative Analysis

The study uses the DecisionTreeRegressor class, RandomForestRegressor class, and GradientBoostingRegressor class in Python's scikit-learn library to construct the decision tree model, random forest model, and GBDT model respectively. As shown in Table 3, the GBDT model performed the best in all indicators, with the smallest MSE and RMSE values and the largest R^2 values, indicating that the GBDT model had the best prediction accuracy and stability, and could better fit the data and predict air quality. The performance of the random forest model is secondary, and the performance of the decision tree model is relatively poor.

Table 3. Performance comparison of different models

Model	MSE	RMSE	MAE	R^2
Decision tree	80.1535	8.9528	3.4190	0.9764
Random forest	60.7539	7.7945	2.6556	0.9821
GBDT	56.9962	7.5496	3.3120	0.9832

3.4.2 GBDT Model Parameter Optimization

The GBDT model, which demonstrated relatively high performance, was further fine-tuned to optimize its parameters. By using the random search method, a relatively excellent parameter combination is searched in a large parameter space. In the random search process, the learning rate is taken in the range of [-2,10], the value range of the tree is [5,200], and the maximum depth is [3,6]. After repeated experiments, when the learning rate is 0.1291549665014884, the number of trees is

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

180, and the maximum depth is 5, the performance of the GBDT model on the test set is further improved, and the index performance is shown in Table 4.

Table 4. Comparison of GBDT model before and after optimization

	MSE	RMSE	MAE	R^2
Before optimization	56.9962	7.5496	3.3120	0.9832
After optimization	49.4692	7.0334	2.4856	0.9854

3.4.3 GBDT Model Prediction and Influencing Factor Analysis

The optimized GBDT model was used to predict the air quality in Shaanxi Province, and the prediction results were obtained. The prediction results are compared with the actual values, and it is found that the predicted values are consistent with the actual values, and some of the prediction results are shown in Table 5.

Table 5. Partial prediction results of AQI index

Date	Actual AQI index	Forecast AQI index	Error
2024/6/9	62	57.961157	4.038843
2024/11/26	63	64.410695	-1.410695
2025/5/27	25	26.085416	-1.085416
2025/6/3	104	109.358313	-5.358313
...	...	...	...

At the same time, the factors affecting air quality prediction are discussed, as shown in Fig.3. By analyzing the importance of features, it was found that PM10 concentration had the greatest impact. In addition to PM10 concentration, the influence of other air quality indicators weakened in turn, ranking PM2.5, O_3, SO_2, CO, and NO_2, which also played a role in air quality prediction to varying degrees.

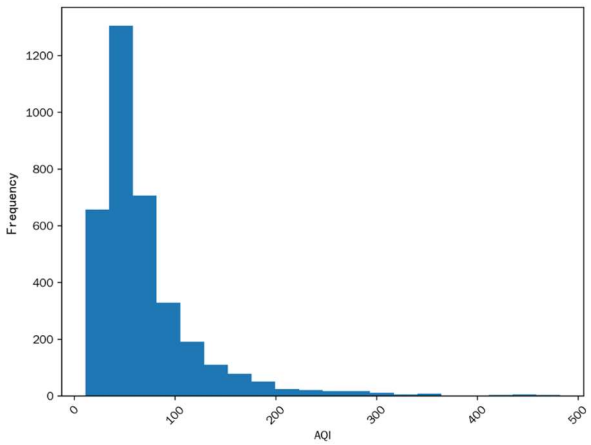

Fig.3 Distribution of feature importance of the GBDT model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0005

Therefore, according to the importance of the influencing factors, in order to ensure air quality, it is necessary to set strict graded emission standards, and formulate targeted emission indicators for industrial enterprises of different sizes and types according to different pollutants, especially PM10, PM2.5 and SO_2. For illegal enterprises, it is necessary to increase the punishment so that the fine is linked to the degree of violation, and if the circumstances are serious, they should be ordered to stop production and rectify. Implement subsidy policies for environmental protection and technological transformation, encourage enterprises to adopt advanced, cleaner production technologies, and provide financial support and tax incentives to enterprises that actively carry out emission reduction and transformation. Encrypt the layout of intelligent air quality monitoring stations, and add stations in key areas such as industrial clusters, traffic arteries, and densely populated areas. Introduce advanced digital intelligence monitoring equipment and conduct regular maintenance and updates to ensure the accuracy and timeliness of monitoring data.

4. Summary

This study focuses on the prediction and application of air quality. It is found that the parameter-optimized GBDT model has the best performance in various evaluation indicators, showing higher accuracy and stability in air quality prediction. At the same time, PM10, PM2.5, and SO_2 are important influencing factors of air quality. This study expands the application practice of machine learning in the field of environmental prediction and verifies the effectiveness of the GBDT model in air quality prediction. The research results can provide a method reference for the development of high-precision environmental prediction systems, improve the level of public health protection response through real-time intelligent early warning, assist environmental supervision decision-making, and provide a technical basis for regional air pollution joint prevention and control.

Acknowledgments

Supported by: Xi'an Eurasia University Research Fund Project "Research on Practical Pathways for Public Collaborative Governance in Environmental Protection Based on Digital Intelligence Technologies" (grant number 2024XJSK04); and 2024 General Special Research Project of Shaanxi Provincial Department of Education "Research on Enhancing the Effective-ness of Public Collaborative Governance in Environmental Protection Empowered by Digital Intelligence Technologies" (grant number 24JK0168).

References

[1] Shaanxi Provincial Market Supervision and Administration Bureau Shaanxi Continues to Deepen Air Pollution Control. Information on: https://snamr.shaanxi.gov.cn/sy/ztzl/dqwrzlzxxd/202508/t20250811_3553300.html

[2] Luo Z, Zeng R, Wang P. Air Quality Prediction Based on Quadratic Prediction Model. Learning & Education. 2022.

[3] Duan J, Gong Y, Luo J, et al. Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. Scientific Reports. 2023, Vol. 13.

[4] Hua V, Nguyen T, Dao M S, et al. The impact of data imputation on air quality prediction problem. PLoS ONE. 2024, Vol. 19 (No. 9), p. 39.

[5] Gao Senhai, Ma Xu. Air pollution level prediction based on Self-Attention CNN-LSTM. Journal of Tianjin University of Technology. 2025, p. 1-7.

[6] Zhou Jianguo, Qin Yuan, Zhou Luming. Air quality index prediction model utilizing quadratic decomposition, LSTM-ELM, and error correction techniques. Journal of Safety and Environment. 2025, Vol. 25 (No. 01), p. 322-334.