Advances in Internet of Things, 2025, 15(4), 87-95
https://www.scirp.org/journal/ait
ISSN Online: 2161-6825
ISSN Print: 2161-6817

Research on Prediction of Air Quality CO Concentration Based on Python Machine Learning

Ziyang Wang

Guangdong Experimental High School, Guangzhou, China
Email: Dannyouxiang@163.com

How to cite this paper: Wang, Z.Y. (2025) Research on Prediction of Air Quality CO Concentration Based on Python Machine Learning. *Advances in Internet of Things*, **15**, 87-95.
https://doi.org/10.4236/ait.2025.154005

Received: October 4, 2025
Accepted: October 25, 2025
Published: October 28, 2025

Abstract

With the accelerating pace of urbanization, the issue of air pollution has become increasingly severe. Notably, carbon monoxide (CO), as a prevalent harmful gas, poses potential threats to both human health and the environment. Therefore, accurate prediction of CO concentration and analysis of its influencing factors are of significant importance for urban environmental management and public health protection. This study utilizes air quality monitoring data from the UCI open database, selecting multidimensional features including gas sensor outputs and meteorological conditions, and employs a Random Forest regression model to predict CO concentrations. By comparing actual values with predicted values, the model's performance was evaluated using Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). The results indicate that the proposed method can, to some extent, accurately reflect the variation trends of CO concentrations. Furthermore, through feature importance analysis, it was found that features such as benzene concentration (C6H6 (GT)), nitrogen oxides (Nox (GT)), nitrogen dioxide sensor readings (PT08.S4 (NO_2)), and carbon monoxide sensor readings (PT08.S1 (CO)) exhibit high contributions in predicting CO concentrations. This research provides a valuable reference for air pollution prediction and intelligent environmental governance.

Keywords

Air Quality Prediction, Carbon Monoxide (CO), Random Forest, Machine Learning, Feature Importance

1. Introduction

In recent years, air pollution has emerged as a critical global environmental and

public health concern. The continuous increase in industrialization and the number of motor vehicles has led to the presence of numerous pollutants in the air, including sulfur dioxide (SO_2), nitrogen oxides (NOx), ozone (O_3), and carbon monoxide (CO). CO, in particular, as a common harmful gas, primarily originates from combustion processes, such as vehicle exhaust, industrial emissions, and residential heating [1]. Excessive inhalation of CO can reduce the oxygen-carrying capacity of hemoglobin in the blood, leading to poisoning and even endangering life [2]. Therefore, accurately predicting changes in CO concentration is crucial for urban environmental management, public health prevention, and air quality improvement.

Traditional air quality prediction methods often rely on physical modeling and chemical reaction mechanism models. While scientifically grounded, these methods often fall short when dealing with complex, non-linear environmental factors [3]. In recent years, with the development of big data and artificial intelligence, machine learning methods have been widely applied in environmental science research. Machine learning techniques, by leveraging historical monitoring data, can uncover complex, non-linear relationships between pollutants and environmental factors, thereby achieving efficient and accurate prediction [4].

This paper selects the Air Quality Data Set [5] from the UCI database, which contains air monitoring data from an Italian city over several consecutive months, including readings from various gas sensors and meteorological conditions. The primary objectives of this study are twofold: (1) to develop a robust predictive model for CO concentration using a Random Forest regressor, and (2) to identify and interpret the key influencing factors through feature importance analysis. The research results can provide references for government and research institutions in air quality monitoring and governance.

2. Data and Methods

2.1. Data Source

The dataset used in this paper comes from the Air Quality Dataset in the UCI Machine Learning Repository. This dataset contains air quality monitoring data from an Italian city between March 2004 and February 2005. The original data comprises 9358 records, with an hourly recording frequency. The data includes:

1) CO (GT): True hourly averaged concentration CO in mg/m^3.

2) PT08.S1 (CO): Tin oxide gas sensor response for CO.

3) NMHC (GT): True hourly averaged overall Non-Methane Hydrocarbons concentration in $microg/m^3$.

4) C6H6 (GT): True hourly averaged Benzene concentration in $microg/m^3$.

5) PT08.S2 (NMHC): Titania gas sensor response for NMHC.

6) PT08.S3 (NOx): Tungsten oxide gas sensor response for Nox.

7) Nox (GT): True hourly averaged NOx concentration in ppb.

8) PT08.S4 (NO_2): Tungsten oxide gas sensor response for NO_2.

9) NO_2 (GT): True hourly averaged NO_2 concentration in $microg/m^3$.

10) PT08.S5 (O_3): Indium oxide gas sensor response for O_3.

11) T: Temperature in ˚C.

12) RH: Relative Humidity in %.

13) AH: Absolute Humidity.

In the original data, some missing values are represented by-200. During preprocessing, these entries were replaced with NaN, and subsequently removed using listwise deletion, resulting in a final dataset of complete instances for analysis.

2.2. Methods and Process

The overall research process of this paper includes six steps: data preprocessing, feature selection, model building, model evaluation, feature importance analysis, and visualization. The specific steps are as follows:

(1) Data Preprocessing

The data used in this study were sourced from the Air Quality dataset. In the original dataset, certain missing monitoring values were denoted by -200. Employing these values directly for modeling would introduce significant interference in the results. Therefore, all instances of -200 were first replaced with NaN. Subsequently, samples containing missing values in the target variable CO(GT) and the selected feature variables were removed to ensure the integrity and reliability of the training data. This process resulted in a cleaner dataset, contributing to enhanced model stability and predictive accuracy. This study employed the listwise deletion method, whereby any sample row containing a missing value in either the features or the target was entirely removed. A potential implication of this method is that if the missing values are not Missing Completely at Random (MCAR), it may introduce sample bias and consequently reduce the model's generalizability. However, in the present dataset, missing values primarily originated from temporary sensor malfunctions. The pattern of missingness is relatively random and the proportion is low (<10%); thus, listwise deletion is not expected to significantly compromise the data's representativeness.

(2) Feature Selection

The original dataset comprises 14 features. Considering the correlation between carbon monoxide (CO) concentration and other gaseous pollutants, sensor signals, and meteorological conditions, this study selected 12 potential influencing factors as input features, excluding the "Time" and "Date" variables. These selected features are: PT08.S1 (CO), NMHC (GT), C6H6 (GT), PT08.S2 (NMHC), Nox (GT), PT08.S3 (NOx), NO2 (GT), PT08.S4 (NO_2), PT08.S5 (O_3), T, RH, and AH. These variables encompass volatile organic compounds (VOCs), nitrogen oxides (NOx, NO_2), ozone (O_3), as well as meteorological factors such as temperature and relative humidity, collectively providing a comprehensive representation of the key characteristics of air pollutant emissions and dispersion. Utilizing this multi-dimensional input can effectively improve the model's capability to capture variations in CO concentration.

(3) Model Establishment

During the modeling phase, this study employed the Random Forest Regressor [6], an ensemble learning method. Random Forest is an algorithm based on decision tree ensembles. By aggregating the predictions (averaging) from multiple trees, it mitigates the variance inherent in a single model, thereby enhancing prediction accuracy and generalizability. Compared to traditional linear regression methods, Random Forest demonstrates greater robustness in handling non-linear relationships and high-dimensional features, making it well-suited for complex environmental monitoring data such as air quality metrics. The model was implemented using the RandomForestRegressor from the scikit-learn library, with the primary hyperparameters configured as follows:

1. n_estimators = 100 (Number of trees in the forest).

2. random_state = 42 (Ensures result reproducibility).

3. Other parameters (e.g., max_depth, min_samples_split) utilized the library's default settings.

(4) Model Evaluation

To quantify the model's prediction performance, this paper uses Mean Absolute Error (MAE) and the Coefficient of Determination (R^2) as evaluation metrics. MAE reflects the average deviation between predicted values and true values; a smaller value indicates predictions are closer to reality. R^2 measures the model's ability to explain the variance in the observed data, ranging from [0,1]. A value closer to 1 indicates a better fit. These metrics provide a standardized and interpretable framework for evaluating the model's predictive performance.

Using these two metrics allows for a comprehensive assessment of the predictive model's performance from both error and goodness-of-fit dimensions.

(5) Feature Importance Analysis

In addition to predicting CO concentration, this paper also uses the built-in feature importance calculation method of Random Forest to analyze the contribution degree of each input variable to the prediction result [7]. This method measures the importance of each feature by counting its average contribution to error reduction during splits in the decision trees. The importance scores are normalized so that they sum to 1, providing a relative measure of each feature's influence. The results show that variables such as C6H6 (GT), PT08.S2 (NMHC), and Nox (GT) have high importance weights, making them the main factors affecting changes in CO concentration. This result not only verifies the important role of VOC and NOx in air pollution mechanisms but also provides data support for subsequent air quality monitoring and control.

(6) Visualization Analysis

To more intuitively display the model's effects, this paper created three types of charts:

a) Scatter plot comparing true values and predicted values: Used to examine the consistency between prediction results and true observations.

b) Bar chart of feature importance: Shows the relative contribution of each input variable to the predictive model, annotating the top five most critical features.

c) Time series prediction trend chart: Taking the first 100 samples of the test set as an example, compares the changing trends of true values and predicted values over time, verifying the model's stability in continuous prediction.

These visualizations serve as critical tools for interpreting the model's performance and outcomes beyond numerical metrics. Through the above steps, this paper implemented a complete research process from data processing and model building to result interpretation. It not only quantitatively evaluated the prediction effect of CO concentration but also identified key pollution factors from a mechanistic perspective, providing methodological and application value for air quality research.

3. Experimental Results and Analysis

3.1. Model Prediction Performance

The dataset was partitioned into a training set and a test set with a ratio of 8:2. Predictions for CO concentration were generated on the test set. Although the data exhibits time-series characteristics, the objective of this study was to explore the static correlations among sensor variables and the model's overall predictive performance, rather than to conduct time-series forecasting. Therefore, random splitting does not introduce look-ahead bias. The resulting evaluation metrics are presented in Figure 1.

1) Mean Absolute Error MAE: approx. 0.179.

2) Coefficient of Determination R^2: approx. 0.968.

Hese values were calculated on the unseen test set, providing an unbiased estimate of the model's generalization performance. The results indicate that the Random Forest model achieved excellent performance in predicting CO concentrations, able to predict CO concentration relatively accurately, and the R^2 close to 1 indicates a good model fit.

Figure 1. Mean absolute error and coefficient of determination.

3.2. Comparison of True Values and Predicted Values

By drawing a scatter plot (Figure 2), it can be observed that most points are distributed along the ideal diagonal line, indicating high consistency between predicted values and true values. Although there are some errors for a few extreme values, the overall trend prediction is relatively reliable.

Furthermore, in the time series comparison of the first 100 samples, the model's predictions closely align with the true observations, effectively reflecting the fluctuation trend of CO concentration.

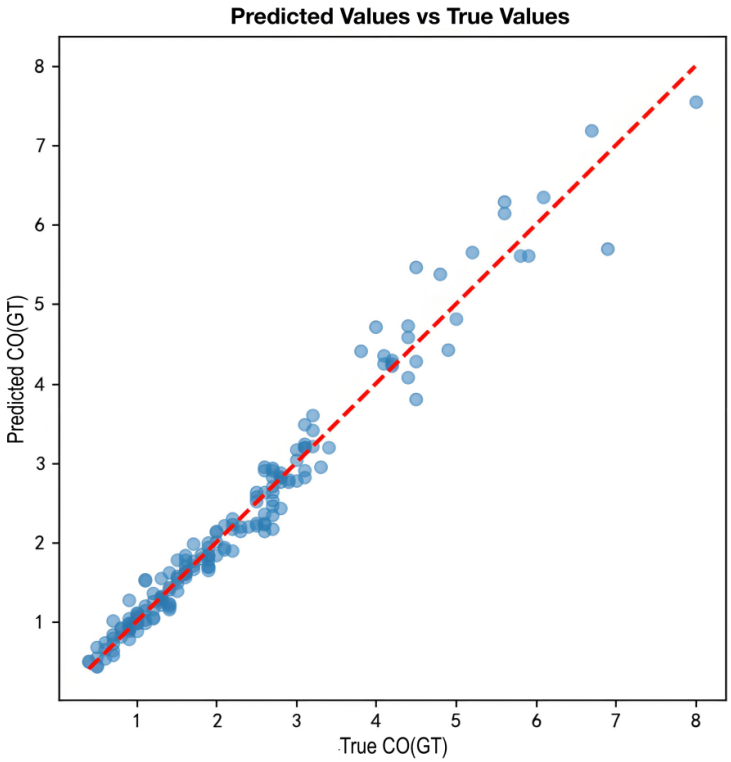

Figure 2. Predicted values vs. true values.

3.3. Point-by-Point Comparative Analysis of the First 100 Samples

To further test the model's fitting ability on short-term sequences, the first 100 samples from the test set (*i.e.*, Figure 3 in the paper) were selected for point-by-point comparative analysis. As shown in Figure 3, the first 100 samples from the test set are selected for comparison, showing the fitting situation of true values and model predicted values on the time series. From the overall trend, the Random Forest regression model can track the dynamic changes of CO concentration, especially in the main range of concentration fluctuations, where the predicted values and true values are basically consistent, showing a strong correlation.

It can be observed in the figure that the predicted values of most data points highly overlap with the true values, indicating the model has strong fitting capability in stable intervals. However, at local extreme points (such as near sample number 42 and sample number 95), the prediction results have certain deviations, often underestimating or overestimating the extreme peaks. This indicates that while the model excels at predicting overall trends, accurately forecasting peak concentrations—critical for public health advisories—remains a challenge, likely due to the underrepresentation of such extreme events in the training data. This may be related to the low proportion of extreme samples in the training data and also reflects the model's shortcomings in coping with sudden pollution events.

Overall, the model's prediction error in the first 100 samples is small, and it can relatively accurately reflect the fluctuation trend of CO concentration, indicating that Random Forest regression has good application potential in air quality prediction.

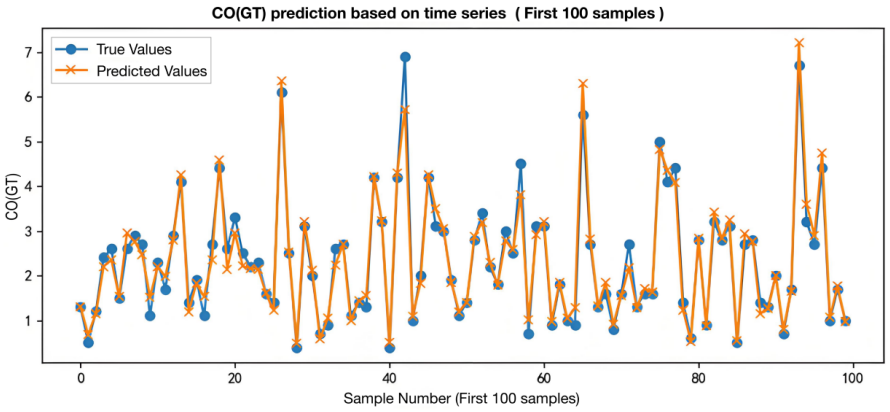

Figure 3. Comparison of the first one hundred samples.

3.4. Feature Importance Analysis

In the feature importance analysis (see Figure 4), the model identified Benzene concentration C6H6 (GT) and the Non-Methane Hydrocarbons sensor PT08.S2 (NMHC) as the two most important influencing factors, with weights reaching 0.459 and 0.411 respectively, accounting for the majority of the contribution. This indicates that Benzene in the air and indicators related to Non-Methane Hydrocarbons are most closely related to CO concentration. Secondly, Nitrogen Oxides Nox (GT), Nitrogen Dioxide sensor PT08.S4 (NO_2), and CO sensor PT08.S1 (CO) also have some influence on the prediction This result aligns with established atmospheric chemistry, wherein benzene and NMHCs are key volatile organic compounds (VOCs) that often share common emission sources (e.g., vehicle exhaust) with CO.

Air pollution mechanisms: Benzene and NMHC are important volatile organic compounds (VOCs), highly correlated with CO emissions, while NOx and NO_2 are often closely related to traffic emissions and also have indicative value for changes in CO concentration.

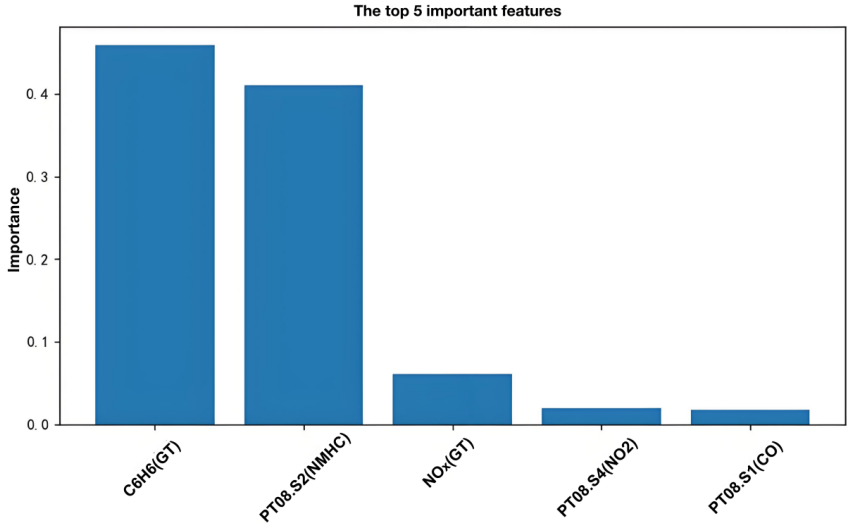

Figure 4. Feature importance ranking.

4. Discussion

This study demonstrates that the machine learning approach can effectively predict CO concentrations and, to a certain extent, elucidate its influencing factors. Compared to traditional physico-chemical models, the Random Forest method does not require the explicit formulation of complex chemical reaction mechanisms and relies solely on the data itself to generate predictions. This characteristic suggests its potential applicability in large-scale, real-time air quality forecasting. Model analysis identified benzene (C6H6 (GT)) as one of the most significant features influencing CO (GT) concentration. This finding holds practical implications: benzene is a major byproduct of vehicle emissions and industrial combustion, and its concentration fluctuations often coincide with increases in carbon monoxide. Consequently, in urban air quality management, the real-time monitoring of benzene levels could serve as a leading indicator for CO pollution, facilitating early warnings for traffic peak hours or industrial emission events, thereby providing a scientific basis for formulating emergency response strategies.

However, this study has several limitations:

The handling of missing values in the dataset was relatively simplistic; future work could explore imputation techniques or deep learning-based methods for missing data completion.

The model selection was confined to Random Forest; subsequent research could extend comparisons to other algorithms such as Gradient Boosting Decision Trees (GBDT) and neural networks.

This study did not account for temporal dependencies; incorporating deep learning models like Long Short-Term Memory (LSTM) networks could better capture time-series characteristics.

5. Conclusion

Based on the UCI Air Quality dataset, this paper used a Random Forest regression model to predict CO concentration and identified the main influencing factors through feature importance analysis. In conclusion, this study successfully developed a Random Forest regression model that achieves high predictive accuracy for ambient CO concentrations and effectively captures their temporal dynamics. The key influential features identified were C6H6 (GT), PT08.S2 (NMHC), and Nox (GT). Future research could combine more machine learning algorithms and time series modeling methods to further improve prediction accuracy and application value. Incorporating real-time traffic and point source emission data could also enhance model performance. The findings of this study offer valuable insights and a practical framework for environmental monitoring, air quality early warning systems, and urban management strategies.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Li, R.X. (2025) Research on Reliability Prediction and Dynamic Maintenance Optimization of Rolling Bearing Based on Random Forest. Master's Thesis, Changchun University of Technology. https://doi.org/10.27805/d.cnki.gccgy.2025.001090

[2] World Health Organization. (2021) WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, SULFUR Dioxide and Carbon Monoxide.

[3] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. and Baklanov, A. (2012) Real-Time Air Quality Forecasting, Part I: History, Techniques, and CURRENT Status. *Atmospheric Environment, 60,* 632-655.

[4] Liu, H.W. (2023) Research on Influencing Factors and Prediction of Urban Air Quality Based on Machine Learning. Master's Thesis, Shandong Normal University. https://doi.org/10.27280/d.cnki.gsdsu.2023.001441

[5] Vito, S. (2008) Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F

[6] Wang, X.Y. (2025) Research on the Impact of Vehicle Emissions on Urban Air Quality Based on Multi-Scale Coupling. Master's Thesis, Shandong Jiaotong University. https://doi.org/10.27864/d.cnki.gsjtd.2025.000070

[7] Yuan, Z.C. (2020) Artificial Intelligence—Analysis of Random Forest Technology. *Technology Innovation and Application, No. 6,* 151-152. https://doi.org/10.19981/j.cn23-1581/g3.2020.06.059