WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijgem.v4n1.72
Journal homepage: https://wepub.org/index.php/IJGEM/index

Air Quality Prediction and Warning Based on Machine Learning

Lin Zhang

School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui, 233030, China
2075709573@qq.com

ABSTRACT

This article explores the factors related to changes in PM2.5 concentration from the perspective of machine learning, predicts daily air quality, and analyzes its warning level. Firstly, construct an indicator system with component factors and climate factors as independent variables, and PM2.5 concentration value as the dependent variable; Next, two machine learning algorithms, linear regression and decision tree regression, were used to construct models for regression prediction. The fitting curve between the predicted values and the true values was used to demonstrate the fitting effect, and it was found that decision tree regression had the best fitting. In two models, we trained predictions with step sizes of 3, 5, 7, and 12, respectively. We called the mean_squared_error standard library in Python to calculate the RMSE for each step, and weighted the RMSE for different step sizes of the two models to obtain the final RMSE. To more accurately predict the PM2.5 concentration value for the required date in the question, we extracted data from the time period of each year, calculated the average of each attribute as the test set, and imported it into the model. We then weighted and summed the predicted values of the two models to obtain the final PM2.5 prediction value. Finally, a visual analysis was conducted on the test set and its prediction results to more intuitively demonstrate the prediction performance.

KEYWORDS

Linear regression; Decision tree regression; Machine learning; Classification; Fitting

1. INTRODUCTION

In today's world, air pollution has become a major environmental challenge on a global scale, with far-reaching impacts on human health, social life and ecosystems. In particular, changes in the concentration of fine particulate matter (PM2.5) are directly related to air quality and its threat to public health. Therefore, accurate prediction of PM2.5 concentration and its impact on air quality index (AQI) is crucial for the development of effective environmental policies and control measures. The aim of this paper is to explore the factors associated with the variation of PM2.5 concentration, build a prediction model, and evaluate the performance of the model through a machine learning approach. The main factors affecting PM2.5 concentration are revealed by technical means such as Pearson correlation coefficient analysis, scatterplot matrix, and random forest classification, and regression prediction is performed using linear regression and decision tree regression algorithms. The ultimate goal is to provide a scientific basis for the government and related departments to help air quality management and improvement efforts.

2. NUMERICAL DISTRIBUTION AND RELATIONSHIP BETWEEN COMPOSITIONAL FACTORS AND CLIMATIC FACTORS INCLUDING ATTRIBUTES

By using a scatter plot matrix, simple visualizations of component factors and climate factors can be performed separately. A scatter plot matrix can display the relationships between multiple variables in a dataset and the data distribution of each variable. We used the scatter_matrix function in the pandas library to draw a scatter plot matrix, and the results are shown in Figure 1.

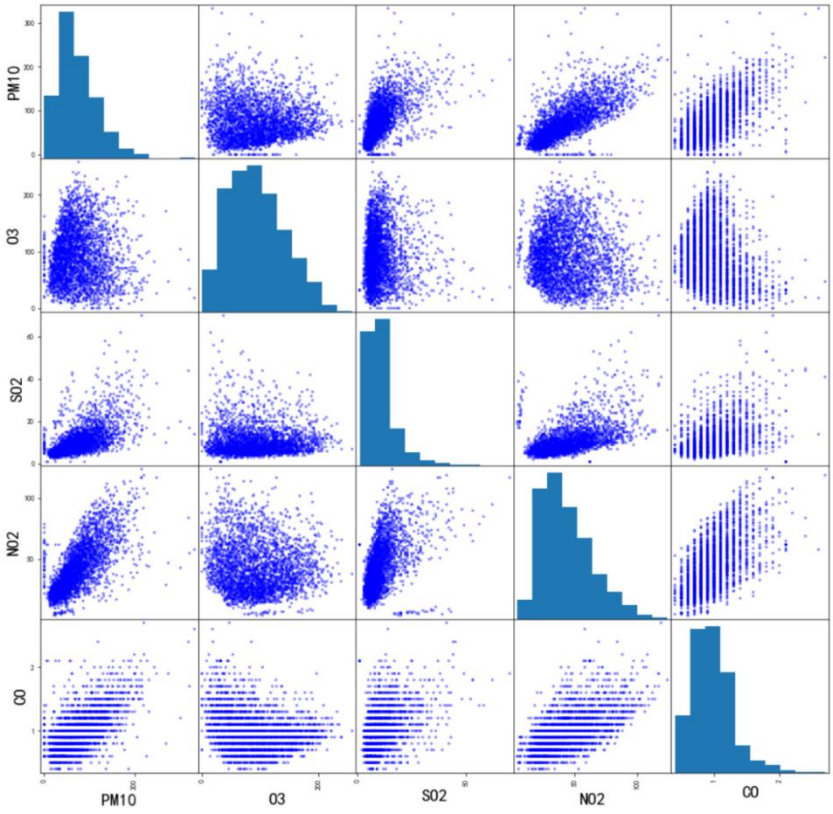

Figure 1. Component Factor Scatterplot Matrix

Random forest classification assesses the extent to which each feature contributes to the classification task by calculating feature importance. This assessment can be based on methods such as mean impurity reduction, replacement importance or Gini importance. By analysing feature importance, we can understand which features have more influence on classification results in random forests, which can help us understand the feature structure of the data and improve model performance.

In sklearn, the decision tree is used to evaluate the importance of features by reducing the amount of Gini purity based on the Gini purity, which of course can be replaced by information gain or information gain ratio. Specifically, for the feature that divides each node in the decision tree, the feature importance is calculated by the formula.

$$importance = \frac{N_t}{N} * (impurity - \frac{N_{tL}}{N_t} * left\_impurity - \frac{N_{tR}}{N_t} * right\_impurity)$$

Where, N is the number of samples; N_t denotes the number of samples in the current node; $impurity$ denotes the purity of the current node; N_{tL} denotes the number of samples in the left child of the current node; $left\_impurity$ denotes the purity of the left child of the current node; N_{tR}

denotes the number of samples in the right child of the current node; $right\_impurity$ denotes the purity of the right child of the current node.

For the factors related to the change of PM2.5 concentration screened by the correlation coefficient heat map, we extracted the feature importance by random forest classification, analysed the degree of influence of compositional and climatic factors on the PM2.5 concentration, and compared and analysed them by bar charts. In sklearn, the feature importance assessment of random forest is mainly calculated based on the feature importance results of multiple decision trees, and the feature importance values obtained from decision trees are also standardised by default, i.e., each dimension is divided by the sum of all dimensions. The results are shown in Tables 1 and 2.

Table 1. Constituent Factor Characteristic Importance

Feature name	Characteristic importance
PM10	72.50 per cent
CO	17.50 per cent
O3	5.90 per cent
NO2	2.60 per cent
SO2	1.50 per cent

Table 2. Importance of climate factor characteristics

Feature name	Characteristic importance
average temperatures	54.10%
measured quantity of rain	13.70 per cent
average pressure	12.60 per cent
Average relative humidity	10.00 per cent
Average 2-minute wind speed	9.60 per cent

Based on the above analyses, we conclude that among the compositional factors, PM10 and CO have the greatest influence on PM2.5 concentrations, followed by O3 and NO2, and SO2 has the least influence on PM2.5 concentrations; among the climatic factors, average air temperature and precipitation have the greatest influence on PM2.5 concentrations, followed by average barometric pressure and average relative humidity, and average 2-minute wind speed has the least influence on PM2.5 concentrations.

3. REGRESSION MODELLING TO PREDICT PM2.5 CONCENTRATION

3.1. Regression Prediction Modelling

Analysing the problem, we can see that the prediction of PM2.5 concentration value is a multi-cause and one-effect problem, with more attributes affecting the result and complex relationships among attributes. Some regression prediction algorithms in machine learning algorithms provide a good idea for solving this kind of problem, so for problem 2, we establish two regression models, Linear Regression and Decision Tree Regression, to predict the concentration of PM2.5.

3.2. Introduction to the Model

3.2.1. Linear Regression

Linear regression is a machine learning algorithm used to establish linear relationships. Its goal is to describe the linear relationship between the independent and dependent variables by fitting a straight line (in the case of one dimension) or a hyperplane (in the case of multiple dimensions). The algorithm

minimises the difference between the predicted values and the actual observed values by finding the best coefficients. Linear regression models can be used to predict continuous dependent variables and are useful for solving regression problems.

For example, given the dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, we try to learn from this dataset to obtain a linear model that reflects the correspondence between $x(i)$ and $y(i)$ as accurately as possible. The linear model, here, is a function of a linear combination of the attributes (x) and can be expressed as:

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

The vector is represented as.

$$f(x) = w^T x + b$$

Where, $w = (w1; w2; w3; \ldots wd)$ denotes the column vector, here w denotes the meaning of weight, which indicates the weight of the corresponding attribute in the prediction result, and is a parameter of the linear model, which is used to calculate the result.

3.2.2. Decision Tree Regression (DTR)

Decision tree regression is a decision tree-based machine learning algorithm for establishing nonlinear relationships between independent and dependent variables. Unlike classification problems, the goal of decision tree regression is to predict continuous dependent variables. Decision tree regression uses a series of decision rules to construct a tree structure where each internal node represents a feature and each leaf node represents a predicted value. During the training phase, the algorithm divides the dataset into subsets by selecting the best features and division points such that the samples in each subset have as similar objective values as possible. This process is recursive until a stopping condition is reached (e.g., maximum depth reached, number of samples less than a threshold, etc.). In the prediction phase, the corresponding leaf node is found through the decision path based on the input independent variables and the predicted value of that node is returned as the output.

Suppose that X and Y are input and output variables, respectively, and Y are continuous variables, given a training dataset of

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

Where, $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots x_i^{(n)})$ is the input instance (feature vector), and$n$ is the number of features, and $i = 1,2, \cdots, N$, and N is the sample capacity.

The feature space is divided using a heuristic method, where each division examines all the values of all the features in the current set one by one, and selects the optimal one of them as the cut-off point according to the squared error minimisation criterion. For example, for the training set No. j feature variable $x^{(j)}$ and its values$s$, as the cut-off variable and cut-off point, and define two regions $R_1(j,s) = \{x|x^{(j)} \leq s\}$ and $R_2(j,s) = \{x|x^{(j)} > s\}$, in order to find the optimal$j$ and$s$, solve the following equation

$$min_{j,s}\left[min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right]$$

That is, find the two regions to be divided so that the sum of the squared errors is minimized j and$s$.

3.3. Model Building and Training

3.3.1. Modelling

The target variable pm25 (i.e., PM2.5 value) was first established, followed by the dependent variable features_1, which stores the feature variables related to PM2.5, including PM10, O3, SO2, NO2, CO, precipitation, average air pressure, average 2-minute wind speed, average air temperature, and average relative humidity.

Create a linear regression model. Linear regression models model the relationship between the independent and dependent variables by fitting a linear function.

Use the Linear Regression class to initialise a linear regression model object, regressor_LR, and then use the fit() method to pass in the feature data and target variables to train the model.

Create a decision tree regression model. A decision tree regression model models the non-linear relationship between the independent and dependent variables by constructing a decision tree. Initialise a decision tree regression model object regressor_DTR using Decision Tree Regress, and again use the fit() method to train the model with feature data and target variables.

3.3.2. Model training

The target variable PM25, and the independent variable feature are divided into training set and test set in the ratio of 7:3, and the training set is imported into the model for training. After that, the test set is imported to get the prediction results, and the comparison curve between the test set and the prediction results is plotted. The comparison of test set and prediction results of the two models is shown in Figure 2 below:

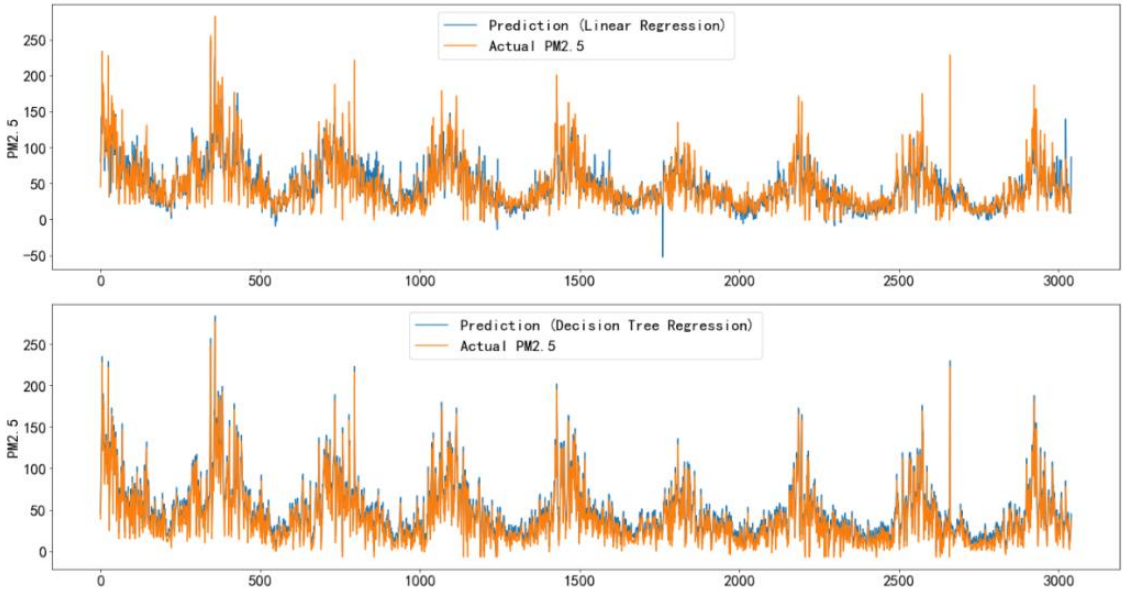

Figure 2. Test set prediction results for both models

From the figure, it can be found that the predicted values obtained from the decision tree regression among the two models are the best fitted to the true values, and the predictions from the linear regression are slightly worse.

For different step lengths, we processed the data in the model accordingly, dividing the data in steps of 3, 5, 7, and 12 for training and prediction, respectively, and obtained the following curves of predicted and true values for different models and different step lengths.

Comparison of PM2.5 predicted by different steps of the linear regression model with the true value is shown in Figure 3:

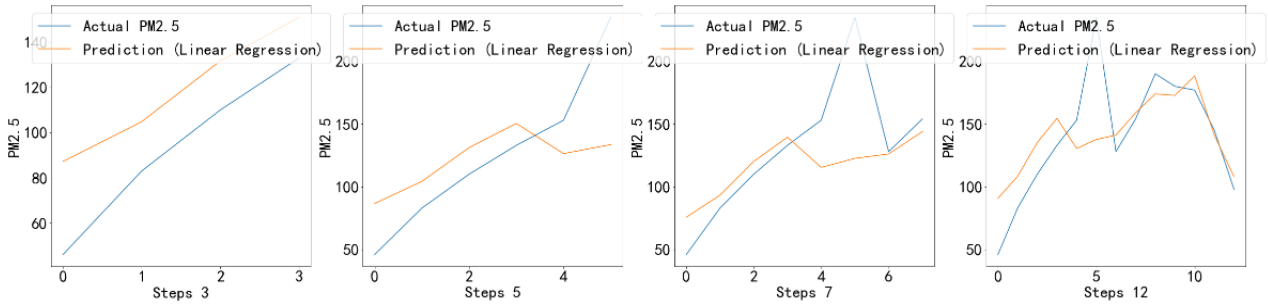

Figure 3. Comparison of predicted and true PM2.5 values for different steps of linear regression

From the figure we can find that in the case of 3 and 5 steps, the predicted values are too far from the true values, but the trend is the same;

In the case of step size 7, 12, the predicted values are close to the true values in most of the cases, in a small area the difference is larger, but the trend is the same. The prediction effect is average.

Comparison of PM2.5 predicted by different steps of the decision tree model with the true value is shown in Fig. 4:

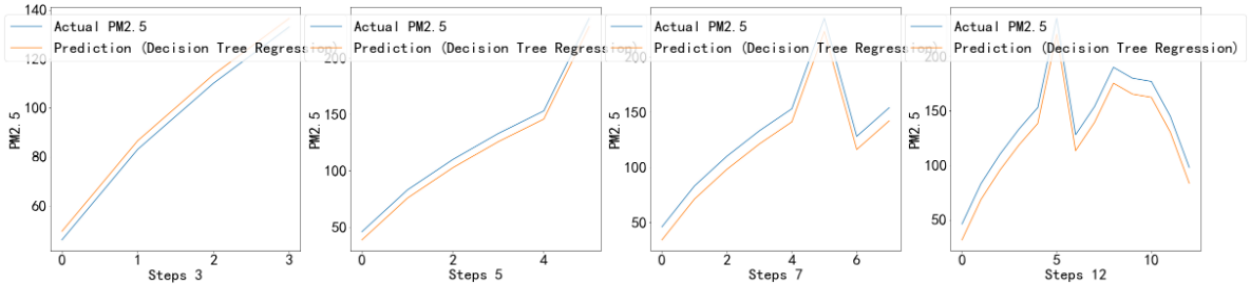

Figure 4. Comparison of predicted and true PM2.5 values at different steps of decision tree regression

From the figure we can find that the predicted values obtained from the decision tree model are very close to the true values and the predictions are extremely good.

4. SOLVING RMSE AND PM2.5

4.1. Solving RMSE

In the prediction process of two different models, RMSE was calculated for each step size of the prediction results. By calling the mean_squared_error standard library in Python, the root mean square error was solved by comparing the test set with the prediction results.

For the linear regression model, the RMSE results obtained at step sizes 3, 5, 7, and 12 are shown in Table 3:

Table 3. Linear Regression Model RMSE Prediction Table

Prediction step size	3-step prediction	5-step prediction	7-step prediction	12-step prediction
RMSE	9.5727	11.8012	10.6276	10.0175

The relatively small RMSE values for each step indicate that the linear regression model has good predictive performance.

For the decision tree model, the RMSE results obtained at step sizes 3, 5, 7, and 12 are shown in Table 4:

Table 4. RMSE Prediction Table of Decision Tree Regression Model

Prediction step size	3-step prediction	5-step prediction	7-step prediction	12-step prediction
RMSE	0.5211	0.7297	0.1040	0.4159

The RMSE values for each step size are very small, indicating that the decision tree model has excellent predictive performance.

4.2. Solving PM2.5

If you want to solve the PM2.5 prediction value for a certain time period through a regression model, you must obtain the numerical values of each attribute for that time period. Therefore, we extract the data segment of this time period from previous years and calculate the average value of each attribute as the test set to import into the model. Obtain the expected values for each date from April 30th to May 11th, 2023, as shown in Table 5:

Table 5. PM2.5 prediction table for solving dates

PM 10	O3	SO2	NO 2	CO	precipitation	Average air pressure	Average 2-minute wind speed	Average temperature	Average relative humidity
84	137.125	10.375	43.75	0.8625	0.460938	1008.259	1.478125	20.12344	72.846875
74.375	121.75	9.375	37.5	0.8625	6.2625	1007.956	1.700625	22.31719	72.9875
70.75	128.625	8.125	33.5	0.8125	1.825	1008.705	1.635	21.9375	73.715625
74.125	134.25	9.375	43.375	0.8375	5.753125	1009.702	1.588125	20.99281	73.73125
78.25	124.25	10.75	45.125	0.8625	0.73125	1009.391	1.55375	22.05469	72.08125
66.375	115	9.25	41	0.8625	2.375	1008.047	1.703125	22.03438	75.7625
53	110.125	9.625	39	0.95	2.009375	1008.038	1.86	22.24063	74.346875
58.875	102.25	8.875	38	0.9375	1.678125	1008.926	3.71875	20.99281	77.621875
51.875	109.375	6.375	33.125	0.8125	8.8	1007.699	1.765625	20.24688	80.696875
67.875	120.125	8.375	41	0.8625	0.405357	1008.703	1.813125	20.35313	78.346875
78.857142	132.1429	10	46.714294	0.914286	2.75	1008.209	2.500714286	21.70786	77.460714
82.33333	103.88889	8.888889	41.55556	0.966667	3.618519	1008.634	1.986111111	20.13472	82.175

Obtain the predicted values as shown in Table 6, and take the predicted average as the final PM2.5 prediction value.

Table 6. Final PM2.5 prediction table for solving dates

Date	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2.5	38.4716	36.5970	35.1949	35.9645	40.4591	33.7573
Date	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2.5	30.9241	33.3049	30.5409	35.9172	45.3066	50.7089

5. CONCLUSION

This article explores the prediction of PM2.5 concentration and analyzes its influencing factors using machine learning methods. Firstly, the main influencing factors were determined through Pearson correlation coefficient and heatmap, and the relationship between component factors and climate factors was demonstrated using a scatter plot matrix. Then, random forest classification was applied to extract feature importance, and linear regression and decision tree regression models were used for prediction. The results show that the decision tree regression model performs the best. Finally, the model was used to predict the test set data and visual analysis was conducted to verify the prediction performance. The research aims to provide scientific basis for policy makers and help improve the level of air quality management.

REFERENCES

[1] Meng Xiaofeng, Hao Xinli, Ma Chaohong et al. Research on Machine Learning Methods in Scientific Discoveries [J]. Chinese Journal of Computer Science, 2023, 46 (05): 877-895.

[2] Xiao Jinjuan, Pang Jinxiang, Chen Wenzhuo. Principal component identification and classification of ancient glass based on random forest model [J]. Science and Technology Innovation, 2023 (14): 37-40.

[3] Ying Xiyuan, Sa Binhan. Prediction of Strawberry Volume and Quality Based on Linear Regression Model [J]. Advanced Mathematics Research, 2023, 26 (03): 86-90.

[4] Zhao Zheng, Yu Xiaojie, Xiong Yuzheng et al. PM2.5 prediction model based on regression analysis and decision tree algorithm [J]. Changjiang Information and Communication, 2022, 35 (11): 9-11.