Nonlinear Dynamics and Systems Theory, **25** (2) (2025) 153–160

Forecasting Air Pollution Levels Using Support Vector Regression and K-Nearest Neighbor Algorithm

I. Indasah [1*], A. Y. P. Asih [2], T. Herlambang [3], P. Triwinanto [4] and K. Oktafianto [5]

[1] *Master of Public Health Department, Universitas STRADA Indonesia, Kediri, Indonesia.*
[2] *Department of Public Health, Faculty of Health, Universitas Nahdlatul Ulama Surabaya, Surabaya, Indonesia.*
[3] *Department of Information System, Universitas Nahdlatul Ulama Surabaya, Indonesia.*
[4] *National Research and Innovation Agency, Indonesia.*
[5] *Department of Mathematics, University of PGRI Ronggolawe, Indonesia.*

Received: October 21, 2024; Revised: March 31, 2025

Abstract: Air pollution is one of the problems faced by big cities, including Surabaya. One of the factors that drives air pollution in big cities is high population mobility. As known, air is composed of oxygen (O2), carbon dioxide (CO2), tropospheric ozone (O3), nitrogen (N2), and particles (PM10 and PM 2.5). High concentration levels of O3 pollutants in an urban area can endanger human health and ecosystems. To monitor air quality in Surabaya city, the city government uses monitoring equipment and air control station facilities. The data obtained becomes a reference for predicting air conditions at that time and forecasting future conditions using certain methods. In this research, the methods used for forecasting are Support Vector Regression (SVR) and K-Nearest Neighbor (K-NN). The Support Vector Regression (SVR) method showed the best error value of 0.0583 and the K-Nearest Neighbor (K-NN) method had the best error value of 0.0486.

Keywords: *pollution; forecasting; Support Vector Regression; K-Nearest Neighbor.*

Mathematics Subject Classification (2020): 62J05, 70-10, 90Bxx.

* Corresponding author: mailto:indasah.strada@gmail.com

1 Introduction

Air pollution is a serious problem faced in big cities, including Surabaya. Some of the contributing factors include high mobility and motorized vehicles. According to the Air Visual's AQI (Air Quality Index) publication in 2019, Surabaya was ranked seventh among the most polluted cities in Indonesia and ranked 226th among the cities of the world [1]. Based on this fact, the Surabaya city government made efforts by issuing Local Regulation Number 3 of 2008 concerning Air Pollution Control of Surabaya City [2].

The air condition of Surabaya with high population mobility is directly proportional to the massive use of motorized vehicles. This is shown by two-wheeled vehicles increasing 7.03% per year from 1,944,802 vehicles in 2015 to 2,081,449 vehicles in 2016 and 2,159,069 in 2017. This resulted in the traffic in the city of Surabaya having a negative impact in the form of air pollution from emissions released by vehicles [3]. Based on these conditions, it is necessary to make a directed and scientific solution for handling such air pollution.

By the development of information technology, it is possible to make a machine learning-based prediction system. The advantages of machine learning are that it provides easier implementation with low computational costs, as well as fast training, validation, testing, and evaluation with high performance compared to physical models, and is relatively less complicated [4].

In this research, we use two methods, namely Support Vector Regression (SVR) and K-Nearest Neighbor (K-NN). Support Vector Regression (SVR) is a development method of the Support Vector Machine (SVM) which is applied for solving regression cases and provides output in the form of continuous data with real numbers so that it can be used for forecasting [5]. K-Nearest Neighbor (K-NN) is a machine learning algorithm used for classification and regression. This algorithm is based on the idea that similar data instances tend to be close to each other in feature space [6], [7].

In previous research, the SVR algorithm was successfully used to forecast the air quality index in Makassar City [8], while the K-NN algorithm was successfully used in predicting air quality in Jakarta City [9]. This research aims to develop and to compare the performance of each method in predicting air pollution levels in the city of Surabaya.

2 Research Methods

The dataset used in this study comes from the Surabaya City air condition data dated 01/01/2020 to 31/12/2020. An overview of the research can be seen in Figure 1 below.

1. **Problem Identification**: This study deals with a case study on the prediction of the air pollution level in the city of Surabaya.

2. **Data Acquisition**: The data used in this study is secondary data sourced from the Surabaya City air conditions as of 01/01/2020 to 31/12/2020.

3. **Data Preprocessing and Analytics**: Basically, data preprocessing and analytics are used to see the initial condition of the dataset obtained from the source. At this stage, the dataset is identified from the data type to missing values that may occur.

4. **Feature Selection**: This study uses the Pearson Product correlation analysis technique to find a linear relationship between two variables having a normal dis-

DATE	PM10	SO2	CO	O3	NO2	MAX	RESULT
01/01/2020	30	20	10	32	9	32	GOOD
02/01/2020	27	22	12	29	8	29	GOOD
03/01/2020	39	22	14	32	10	39	GOOD
04/01/2020	34	22	14	38	10	38	GOOD
05/01/2020	35	22	12	31	9	35	GOOD
06/01/2020	46	23	16	32	9	46	GOOD
07/01/2020	37	23	26	33	11	37	GOOD
08/01/2020	41	26	20	30	11	41	GOOD
09/01/2020	52	23	29	24	12	52	AVERAGE
10/01/2020	24	24	18	25	8	25	GOOD
11/01/2020	34	31	25	23	8	34	GOOD
12/01/2020	27	23	9	33	4	33	GOOD
13/01/2020	33	26	12	36	8	36	GOOD
14/01/2020	34	28	13	27	7	34	GOOD
15/01/2020	29	22	13	36	8	36	GOOD
01/01/2020	30	20	10	32	9	32	GOOD
16/01/2020	52	60	19	30	8	60	AVERAGE
...	...	...	...	...	...	...	...
31/12/2020	18	13	6	24	3	24	GOOD

Table 1: Dataset.

tribution [10]. Below is the function of the Pearson Product:

$$r_{xy} = \frac{N\Sigma XY - (X)(Y)}{\sqrt{N\Sigma X^2 - \Sigma X^2 N\Sigma Y^2 - \Sigma Y^2}}. \tag{1}$$

Notes: r_{xy} is the relationship coefficient; N is the number of samples used; X is the total score of questions; Y is the sum of total scores.

5. **Model Selection**: Support Vector Regression (SVR) is a development algorithm of the Support Vector Machine (SVM) algorithm introduced by Cortes and Vapnik [11]. Like SVM, SVR also uses the best hyperplane in the form of a regression function by making the error as small as possible. The function of the SVR can generally be written as follows:

$$f(x) = w\varphi(x) + b \tag{2}$$

with $f(x)$ being the regression function; w being the vector; b being the bias and the decision boundary equation

$$\begin{aligned} W_x + b &= +e \\ W_x + b &= -e \end{aligned}$$

so that the hyperplane fulfills the inequality

$$e < y - W_x + b < +e$$

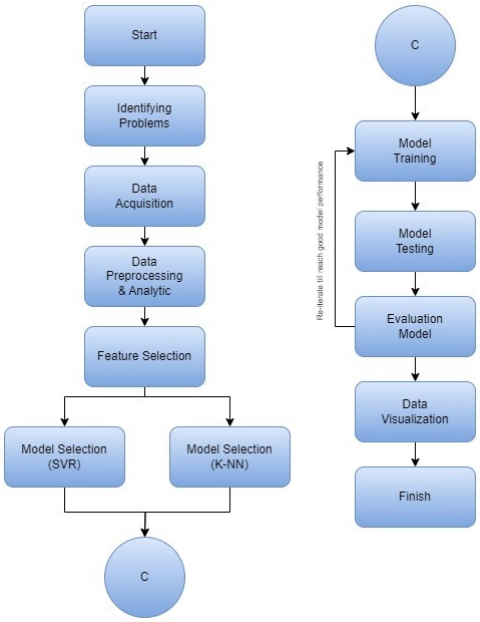

Figure 1: Research methods.

with the minimization function

$$MIN\ \frac{1}{2}\|w^2\| + C\sum_{i=1}^{n}|\xi_i|$$

and the constraint function

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i|.$$

Then, for the K-Nearest Neighbor (K-NN) algorithm calculated using the Euclidean rule, the mathematical function can be written as follows:

$$D = \sqrt{(x_2 - x_1) * 2 - (y_2 - y_1) * 2} \tag{3}$$

with D being the distance; x being the data sample; y being the testing data.

6. **Model Training**: At this stage, the predicted values of the Support Vector Regression (SVR) and K-Nearest Neighbor (K-NN) algorithms are trained based on the division of training data and testing data to obtain error values and accuracy values.

7. **Model Testing**: Model testing of the learning outcomes with the prepared test data.

8. **Evaluation Model**: At the evaluation stage, the model trained and tested is calculated for accuracy based on the resulting error value. This study uses the

Root Mean Square Error (RMSE) method to calculate the error value generated by the model. The function of the Root Mean Square Error (RMSE) is as follows:

$$RMSE = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}} \tag{4}$$

with n being the quantity of data; y_i being the actual value at the i-th data; $\hat{y}_i$ being the predicted value at the i-th data.

3 Result and Discussion

This study tries to implement two machine learning algorithms, namely SVR and KNN, to forecast air pollution using the Python programming language. This study shows the comparison based on the methods and the difference in the quantity of training data and testing data shown in Figures 2-4. Figure 2 is the simulation result of the SVR and K-NN algorithms with 70% of training data and 30% of testing data.

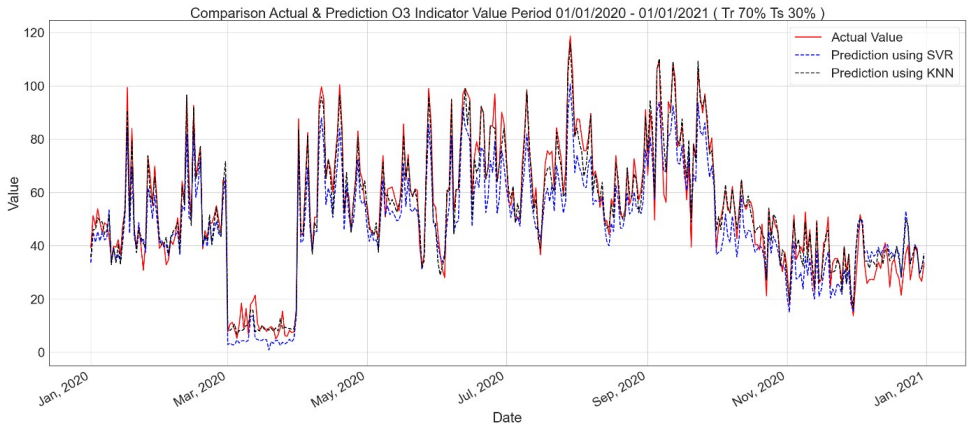

Figure 2: Results of forecasting air pollution by 70% of training data and 30% of testing data.

The simulation results in Figure 2 show the performance comparison of the SVR and K-NN methods in predicting O3 levels in Surabaya with a data split into 70% of training data and 30% of testing data. It appears in the simulation results in Figure 2, that the simulation results using the KNN have a smaller error than those of the SVR, with the RMSE of the KNN of around 0.0486 and the RMSE of the SVR of around 0.0583.

It can be seen that the prediction using the KNN (marked by the black line) is more accurate and consistently close to the actual value (marked by the red line) compared to the prediction using the SVR (marked by the blue line). The graph clearly illustrates that the KNN is more effective in following the fluctuations and dynamics of O3 historical data, thus providing predictions that are closer to reality.

The simulation results in Figure 3 show the performance comparison of the SVR and K-NN methods in predicting O3 levels in Surabaya with a data split into 80% of training data and 20% of testing data. It appears in the simulation results in Figure 3 that the simulation results using the KNN have a smaller error than those of the SVR, with the RMSE of the KNN of around 0.0504 and the RMSE of the SVR of around 0.0588.

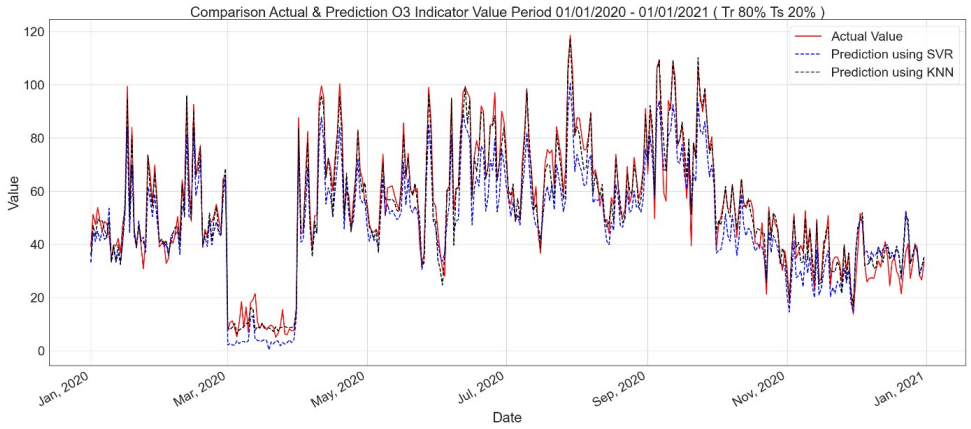

Figure 3: Results of forecasting air pollution by 80% of training data and 20% of testing data.

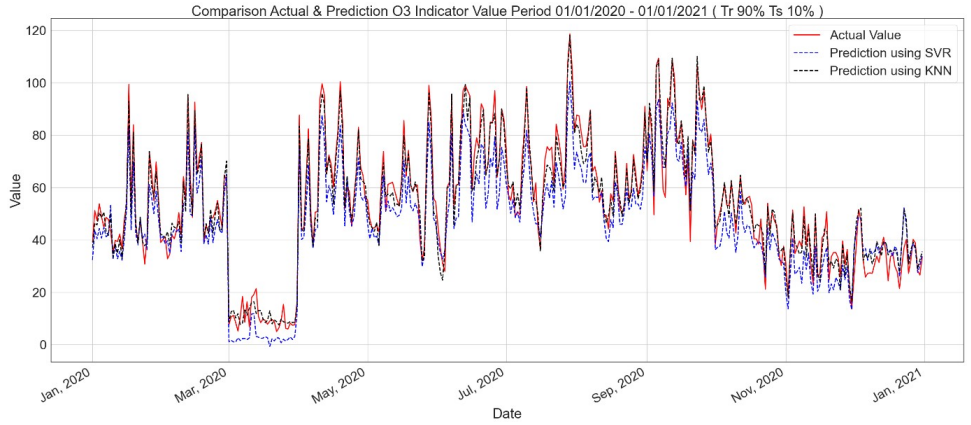

Figure 4: Results of forecasting air pollution by 90% of training data and 10% of testing data.

The simulation results in Figure 4 show the performance comparison of the SVR and K-NN methods in predicting O3 levels in Surabaya with a data split into 90% of training data and 10% of testing data. It appears in the simulation results in Figure 4 that the simulation results using the KNN have a smaller error than those of the SVR, with the RMS of the KNN method having a smaller error than that of the SVR method. A recapitulation of the results of each simulation can be seen in Tables 2-4.

In Table 2, it can be seen that the RMSE value produced by the SVR algorithm is the best in the first simulation with a ratio of 70% of training data and 30% of testing data. Then, in the second and third simulations, there is an increase in the RMSE value with a difference of 0.0005 to 0.0046 compared to the first simulation.

In Table 3, it can be seen that the RMSE value produced by the KNN algorithm is best in the first simulation with a ratio of 70% of training data and 30% of testing data. Then, in the second and third simulations, there is an increase in the RMSE value, which has a difference of 0.0014 to 0.0018 compared to the first simulation. A complete

	70% Training Data and 30% Testing Data	80% Training Data and 20% Testing Data	90% Training Data and 10% Testing Data
RMSE from Forecasting Results	0.0583	0.0588	0.0629

Table 2: Comparison of SVR RMSE Values

	70% Training Data and 30% Testing Data	80% Training Data and 20% Testing Data	90% Training Data and 10% Testing Data
RMSE from Forecasting Results	0.0486	0.0504	0.0500

Table 3: Comparison of SVR-KNN RMSE values.

comparison of the RMSE values of the two algorithms can be seen in Table 4.

	70% Training Data and 30% Testing Data		80% Training Data and 20% Testing Data		90% Training Data and 10% Testing Data	
	SVR	KNN	SVR	KNN	SVR	KNN
RMSE from Forecasting Results	0.0583	0.0486	0.0588	0.0504	0.0629	0.0500

Table 4: Comparison of KNN RMSE values.

In Table 4, it can be seen that the RMSE value produced by the SVR and KNN algorithms is the best in the first simulation with a ratio of 70% of training data and 30% of testing data. Then, in the second and third simulations, there is an increase in the RMSE value occurring in each algorithm. The table above shows that the RMSE value of the simulation results generated by the two algorithms does not touch 1%.

4 Conclusion

Based on the simulation results obtained, it can be concluded that the first simulation results of the SVR and KNN algorithms are the best simulation results with an SVR RMSE value of 0.0583 and a KNN RMSE value of 0.0486. These results prove that the SVR and KNN methods provide good prediction results.

References

[1] M. M. Syaifulloh. Prediction of Air Pollution Standard Index in Surabaya City Based on Carbon Monoxide Gas Concentration. *Jambura Journal of Probability and Statistics* **2** (2) (2021) 86–95.

[2] M. R. Fithori, M. N. Ubaidillah and M. Z. A. Mukminin. Air Pollution Control Through Regional Regulations. *Ma'mal: Journal of Sharia and Law Laboratory* **5** (1) (2024) 73–94.

[3] G. A. Setyo and R. E. Handriyono. Analysis of the Distribution of Carbon Monoxide (CO) Gas from Transportation Sources on Kertajaya Indah Highway, Surabaya. *Environmental Engineering Journal ITATS* **1** (1) (2021) 18–26.

[4] N. M. O. W. S. Sari, H. Elindra and A. H. Saputra. Carbon Monoxide Prediction Using Machine Learning Model Based on Comparison of Time Series Models Case Study of DKI Jakarta. *Collaborative Journal of Science* **7** (3) (2024) 1116–1128.

[5] M. E. S. Wicaksono, G. M. A. Sasmita and I. P. A. E. Pratama. Air Quality Forecasting in Central Jakarta City Using Long Short-Term Memory and Support-Vector Regression Methods. *JITTER - Scientific Journal of Technology and Computers* **4** (1) (2023).

[6] I. Irwansyah and A. D. Wiranata. Comparison of Decision Tree Algorithms, Naive Bayes and K-Nearest Neighbor to Determine Air Quality in DKI Jakarta Province. *Infotech: Journal of Technology Information* **9** (2) (2023) 193–198.

[7] T. Herlambang, V. Asy'ari, R. P. Rahayu, A. A. Firdaus and N. Juniarta. Comparison of Naive Bayes and K-Nearest Neighbor Models for Identifying the Highest Prevalence of Stunting Cases in East Java. *BAREKENG: Journal of Mathematics and Its Applications* **18** (4) (2024) 2153–2164.

[8] R. W. Rahmat, S. Annas and Z. Rais. Support Vector Regression (SVR) Analysis to Predict Air Quality Index in Makassar City. *VARIANSI: Journal of Statistics and Its application on Teaching and Research* **5** (3) (2023) 104–117.

[9] A. Amalia, A. Zaidiah and I. N. Isnainiyah. Air Quality Prediction Using K-Nearest Neighbor Algorithm. *JIPI (Scientific Journal of Informatics Research and Learning)* **7** (2) (2022) 496–507.

[10] R. M. Tedja, M. Arifin and E. S. Agustian. Correlation Analysis of Airbus A320-200 Aircraft Age Against the Amount of Corrosion Occurring Using the Pearson Product Moment Correlation Method. *JIPI (Journal of Aerospace Technology)* **8** (2) (2023).

[11] R. Puspita, H. Cipta and R. Aprilia. Application of the Support Vector Regression Method with the Grid Search Algorithm to Predict Movement Gold Price. *Science Incandescent Journal* **19** (2) (2024) 380–385.

[12] V. Asy'ari, M. Y. Anshori, T. Herlambang, I. W. Farid, D. F. Karya, and M. Adinugroho. Forecasting average room rate using k-nearest neighbor at Hotel S. In: *International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, IEEE, Nov. 2023, 496-500.

[13] D. Novita, T. Herlambang, V. Asy'ari, A. Alimudin, and H. Arof. Comparison Of K-Nearest Neighbor And Neural Network For Prediction International Visitor In East Java. In: *BAREKENG: Journal of Mathematics and Applied Sciences* **18** (3) (2024) 2057–2070.

[14] M. Y. Anshori, V. Asy'ari, T. Herlambang, and I. W. Farid. Forecasting occupancy rate using neural network at Hotel R. In: *International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, IEEE, Nov. 2023, 347-351.

[15] F. A. Susanto, M. Y. Anshori, D. Rahmalia, K. Oktafianto, D. Adzkiya, P. Katias and T. Herlambang. Estimation of Closed Hotels and Restaurants in Jakarta as Impact of Corona Virus Disease (Covid-19) Spread Using Backpropagation Neural Network. In: *Nonlinear Dynamics and Systems Theory: An International Journal of Research and Surveys* **22** (4) (2022) 457–467.

[16] M. Y. Anshori, T. Shawyun, D. V. Madrigal, D. Rahmalia, F. A. Susanto, T. Herlambang, and D. Adzkiya. Estimation of closed hotels and restaurants in Jakarta as impact of corona virus disease spread using adaptive neuro fuzzy inference system. In: *IAES International Journal of Artificial Intelligence (IJ-AI)* **11** (2) (2022) 462–472.