Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME

Tiara Melati Putri Wiryawanto [(1)], Zuyyina Hawani [(2),] Muhammad Attar Ramadhani [(3)]

[1,2,3] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Sunan Ampel Surabaya

Jalan Ir. H. Soekarno No. 682, Gunung Anyar, Surabaya

e-mail: tiarawir@gmail.com[(1)], zuyyinahawani@gmail.com[(2)], attarramadhani@gmail.com[(3)]

ABSTRAK

Udara merupakan komponen terpenting bagi makhluk hidup di bumi. Namun, perubahan-perubahan yang ada di bumi menyebabkan permasalahan yang salah satunya pencemaran udara. Aktivitas manusia menjadi salah satu penyebab pencemaran udara. Hal inilah yang menjadikan kualitas udara masa depan penting untuk diprediksi. Untuk melakukan prediksi kualitas udara, penelitian ini menggunakan metode Support Vector Machine (SVM) dan Autoregressive Integrated Moving Average (ARIMA). Untuk SVM sendiri mempresentasikan salah satu metode dari teknik machine learning. Sedangkan ARIMA mempresentasikan salah satu metode dari model statistik. Dengan menggunakan data dari situs Open Data Jakarta mengenai pengukuran Indeks Standar Pencemaran udara (ISPU) pada lima stasiun pemantau kualitas udara (SPKU) di Provinsi DKI Jakarta tahun 2021, kemudian dilakukan analisis untuk membandingkan performa dan keakuratan dua metode ini dalam memprediksi kualitas udara. Hasil dari penelitian ini menunjukkan bahwa antara pengujian ARIMA dan SVM, dapat dikatakan pengujian SVM memiliki hasil akurasi yang lebih tinggi. Hal ini dapat dilihat dari rata-rata hasil akurasi dengan beberapa perlakuan.
Kata kunci: Kualitas Udara, Support Vector Machine, ARIMA

ABSTRACT

Air is the most important component for living things on earth. However, the changes that exist on earth cause problems, one of which is air pollution. Human activity is one of the causes of air pollution. This is what makes future air quality feasible to predict. To predict air quality, this research use the Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) methods. For SVM itself, it presents one of the methods of machine learning techniques. Meanwhile, ARIMA presents one of the methods of the statistical model. Using data from the Open Data Jakarta website regarding measurements of the Air Pollution Standard Index (ISPU) at five air quality monitoring stations (SPKU) in DKI Jakarta Province in 2021, an analysis was then carried out to compare the performance and accuracy of these two methods in predicting air quality. The results of this study indicate that between ARIMA and SVM testing, it can be said that SVM testing has higher accuracy results. This can be seen from the average accuracy results with several treatments.
Keywords: *Air Quality, Support Vector Machine, ARIMA*

INTRODUCTION

Over the last few decades, most of humanity transformed into city dwellers. However, this rapid change towards urbanization gave rise to several problems, including air pollution [1]. Air is one of the many components that are important in everyday life for living things on earth. Air is a mixture of gases, which is composed of 78 percent Nitrogen, 20 percent Oxygen, 0.3 percent Carbon Dioxide (CO2) and the remainder consists of Neon (Ne), Helium (He), Methane (CH4) and Hydrogen (H2) [2].

Air can be said to be polluted or polluted when it gets additional other gases that can cause disturbances and changes in composition. Air pollution results from a combination of high emissions and unfavorable weather [1]. Air pollution itself consists of a mixture of particulate matter (PM2.5 and PM10) and gas species (NO2, CO, O3, and SO2), which have acute and chronic effects on human health, especially among adolescents and the elderly. [3]. Emissions of air pollutants and their precursors determine regional air quality and can change climate [4]. Entering an era of rapidly changing climate, the impact on air quality needs to be better understood and reviewed. This is good for air quality management purposes and as one of the social impacts of climate change.

Human activity is one of the causes of air pollution. Air pollution is a process of entering substances, energy, and other components into the air which causes air quality to decrease and function improperly. [2]. This is why, in addition to monitoring, there is a demand to predict air quality in the future. One of the aims is to be able to inform the government or authorities in policy making, such as conducting traffic control when the air is heavily polluted.

One of the tools for prediction or classification is to use the support vector machine (SVM) method. SVM is one of the classification methods with a fairly high degree of accuracy in making predictions on air quality assessments. SVM itself has several advantages in classifying a pattern accurately even though the dataset has limitations. On the other hand, SVM has limitations when the number of attributes used tends to be large which results in heavy computational loads and reduced accuracy [5].

Several studies related to air quality prediction have been carried out in the last decade. Lu et al., in 2002 predicted air pollutant parameters by comparing the SVM and the classical radial basis function (RBF) network [6]. The results of this study indicate that SVM is superior to conventional RBF in predicting air quality parameters with different time series. Then, research related to air pollution prediction using the SVM method was developed by Dun et al., in 2020 which predicted short-term air quality using a hybrid model, which combines the fractional gray linear regression and SVM methods [7]. The results of this study indicate that the hybrid model is used to predict three air pollutants (PM10, PM2.5, and NO2) in Shijiazhuang and Chongqing, and it appears that the prediction accuracy of the hybrid model is significantly higher than that of the single model.

Besides SVM, a method that can be used to make predictions is ARIMA. This method has several stages including identification, estimation, diagnostic check and finally forecasting. The models in ARIMA itself are divided into three, namely the Autoregressive Model, the Moving Average, and the Mixed Model which have the characteristics of the two previous models. However, before carrying out these stages the data must be ascertained to be stationary first. That

is, the data does not increase or decrease. This is because the requirements for the formation of the ARIMA model are stationary data.

Research on air pollution prediction using the ARIMA method shows that the ARIMA model is suitable for short-term predictions if the data is stationary [8]. Subsequent research related to air quality prediction by comparing NO2 and SO2 using the ARIMA method [9]. The results of this study indicate that the ARIMA model is the most effective model for evaluating results through analysis and prediction. With this method it basically helps environmental technicians to study and analyze air pollution levels, and therefore helps the government to take preventive actions.

From these previous studies, in this study a comparison will be made of machine learning techniques using the support vector machine (SVM) method and statistical models using the autoregressive integrated moving average (ARIMA) method. These two methods are used to compare which method is more suitable for use in air quality datasets. Currently, there is a lot of information available on the internet about air quality in several places. In this study, a sample dataset was taken from the Open Data Jakarta website regarding air quality information for DKI Jakarta in 2021. From this dataset, the performance and accuracy of these two methods in predicting air quality will be compared.

METHOD

The research uses two methods, namely SVM and ARIMA. The SVM method itself represents one of the methods of machine learning techniques. Meanwhile ARIMA represents one of the methods of the statistical model. Actually, there are many other methods for making predictions both from statistical models and machine learning techniques. However, these two methods are considered to have significant advantages when used based on several previous studies.

In comparing the two methods, namely SVM and ARIMA to predict air quality, a flow design is needed to provide a reference or direction in conducting this research. The flow of this research consists of six steps that are done sequentially. The following in Figure 1 is a research flow design.

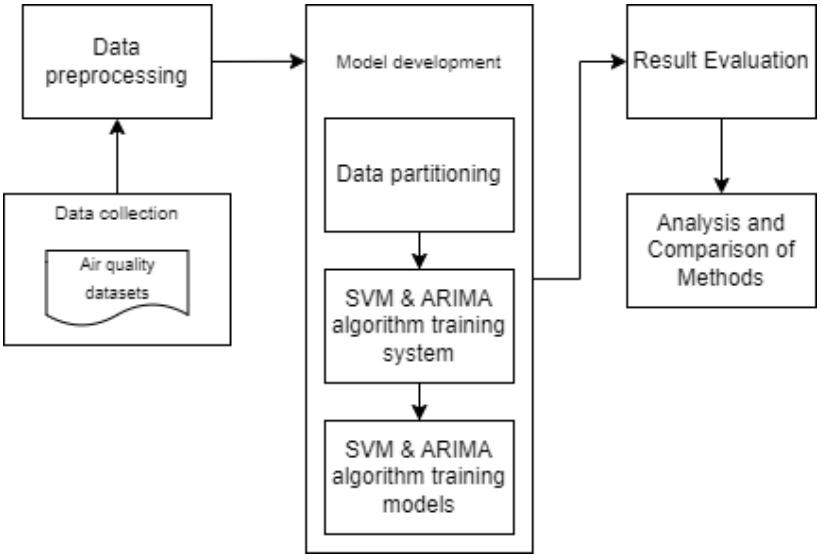

Figure 1. Research flow design

A. Data Collection

The data source used in this study is secondary data obtained from the Open Data Jakarta website regarding measurements of the Air Pollution Standard Index (ISPU) at five air quality monitoring stations (SPKU) in DKI Jakarta Province in 2021. From this dataset, one variable is then defined response or dependent and seven predictor or independent variables as shown in Table 1 below.

Table 1. Variabel dan indikator penelitian

Attribute	Indicator Variables
Y	Category
X1	PM_{10}
X2	$PM_{2.5}$
X3	SO_2
X4	CO
X5	O_3
X6	NO_2
X7	Max

B. Data Preprocessing

At this stage, the data that has been obtained before will be processed in stages which consist of several steps, including:

1. Data cleaning is performed on data that is null and irrelevant to the purpose of analysis.
2. Data transformation is performed, which is used when data needs to be converted to suit the purpose of analysis.
3. Normalization is carried out so that the data used in the two methods does not have large deviations. Analysis using Python with StandardScaler library.

C. Model Development

At the development stage of this model consists of several sub-steps. First, after the data has been processed and modeled, data partitioning is performed. This stage divides the data by dividing the data for training and testing in four experiments, namely with a ratio of 80:20, 70:30 and 60:40. Then from the training data, the algorithms of the two methods are implemented, namely SVM and ARIMA. Then, testing will be carried out, namely predicting the data.

D. Result Evaluation

The results of the data that have been predicted will be tested by model testing and assessing forecasting performance with indicators of percentage accuracy, RMSE and MAE.

1. RMSE

   Root Mean Square Error (RMSE) is a method for measuring the bias or difference in the prediction value of the model.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{I=1}^{n}(Yi - \underline{Yi})^2}{n}} \qquad (1)$$

2. MAE

   Mean Absolute Error (MAE) is the average value of the absolute difference between the actual (actual) value and the predicted (forecasting) value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \text{Ai} - \text{Fi} | \qquad (2)$$

E. Analysis and Comparison of Methods

From the evaluation results of RMSE and MAE for each method using several kinds of partitioning methods, then analysis and comparison are carried out to see the advantages and disadvantages of each method.

F. Support Vector Machine (SVM)

In machine learning, SVM is a supervised learning model that is usually used for pattern recognition, classification, and regression analysis [10]. SVM works based on the Structural Risk Minimization (SRM) principle to find the best hyperplane that separates the two classes in the input space. The performance and effectiveness of SVM itself is greatly affected by the type of kernel functions implemented. The use of the kernel aims to transform data into a high-dimensional space, namely by making non-linear data separate linearly [11]. There are several types of kernel functions that can be implemented based on data criteria. This study uses the two most popular and frequently used kernel functions, namely RBF (Radial Basis Function) and Polynomial.

a. RBF (Radial Basis Function)

   Gaussian radial basis function (RBF) is one of the most widely used kernel functions. RBF itself is usually used when the data is not linearly separated. The following is the kernel's RBF equivalent.

$$K(x,z) = exp\left[-\gamma\left||x-z|\right|^2\right] \qquad (3)$$

b. Polynomial

   The polynomial kernel function is a kernel function that is used when the data is not linearly separated. This kernel function is appropriate when the training dataset has been normalized. The following is the kernel polynomial equation.

$$K(x,z) = (x^T z)^d \; atau \; (1 + x^T z)^d \qquad (4)$$

With x^T is the training data, z is the testing data, and d is the degree of the polynomial

G. Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model dominates in the field of time series forecasting. ARIMA stands for AutoRegressive (AR), Integrated (I), Moving Average (MA). Each of these phrases describes a different part of the mathematical model. Here are the equations ARIMA$(p,d,q)(P,D,Q)$ where (p,d,q) is the non-seasonal part of the model and (P,D,Q) adalah bagian *seasonal* dari model is the seasonal part of the model.

$$\phi_p(B)\Phi_P(B^x)\nabla^d\nabla_s^D Y_t = \theta(B)_Q(B^s)_{e_t} \quad (5)$$

RESULT AND DISCUSSION

The dataset used in this study was published in June, 2022. This dataset consists of 365 data with several attributes related to components in the air. Some of the attributes in the dataset are shown in Table 2 below.

Table 2. Attributes used in research

No	Attribute	Description	Data Type
1	Date	Air quality measurement date	datetime
2	PM_{10}	Particulate matter	integer
3	$PM_{2.5}$	Particulate matter	integer
4	SO_2	Sulfide (in SO_2 form)	integer
5	CO	Carbon monoxide	integer
6	O_3	Ozone	integer
7	NO_2	Nitrogen dioxide	integer
8	Max	The maximum measured value of all parameters measured simultaneously	integer
9	Critical	The maximum measured value of all parameters measured simultaneously	string / object
10	Category	The air quality category is based on the calculation of the air pollution standard index	string / object

From the dataset, data cleaning is then carried out by removing the Critical column because it is irrelevant to the predicted results and the data type does not match the method to be used. After cleaning the data, data transformation is then performed to change the Category data type to integer. This is done so that ARIMA reads the response variable and can make predictions about the Category.

At the last data preprocessing stage, namely normalization. Python uses the StandardScaler library and KNIME uses the Normalizer library. After passing through the data preprocessing stage, data partitioning is carried out using 3 scenarios, namely with a ratio of 80:20, 70:30 and 60:40. From the data partition, then each method is implemented with two tools, namely Python and KNIME. Figure 2 illustrates the flow of SVM and ARIMA modeling using KNIME. The nodes used in this workflow will also be explained in Table 3.

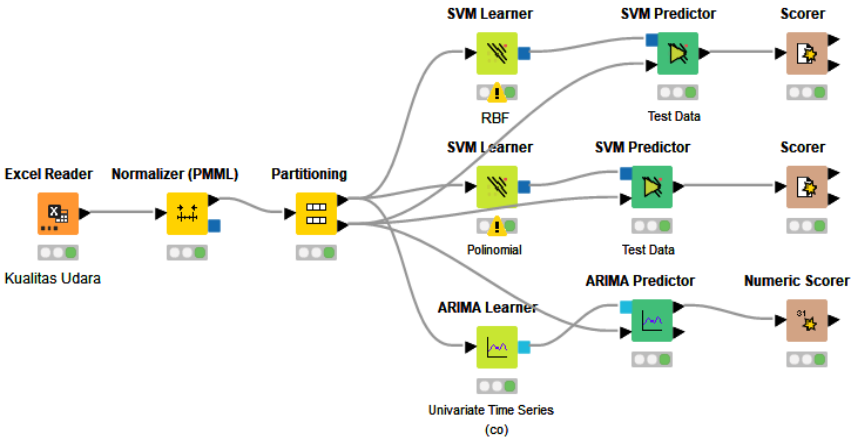

Figure 2. Workflow of SVM and ARIMA using KNIME

Figure 2 above shows the workflow in SVM and ARIMA modeling using KNIME. From the several nodes used [12], a description of each node will be explained in Table 3 below.

Table 3. Nodes used in SVM

No	Node	Description
1	Excel Reader	Reading excel files (xlsx, xlsm, xlsb, and xls formats)
2	Normalizer (PMML)	Normalizes all numeric columns
3	Partitioning	Split data by row to train and test data
4	SVM Learner	Forward partitioned input data using the kernel and parameters in the form of HyperTangent, Polynomial, and RBF
5	ARIMA Learner	Estimating time series parameters on the ARIMA model
6	ARIMA Predictor	Calculating predictions from estimated ARIMA models (forecast and In-sample production)
7	SVM Predictor	Predicting the output value issued by the SVM learner
8	Numeric Scorer	Calculates certain statistics from the predicted results
9	Scorer	Shows the results of the comparison of attribute columns and matrices

A. Support Machine Vector (SVM) Experiment Results

Testing the method to be carried out in this study is based on the distribution of training data and testing using the dataset division procedure. Figure 3 shows the results of data classification based on air quality after training data.

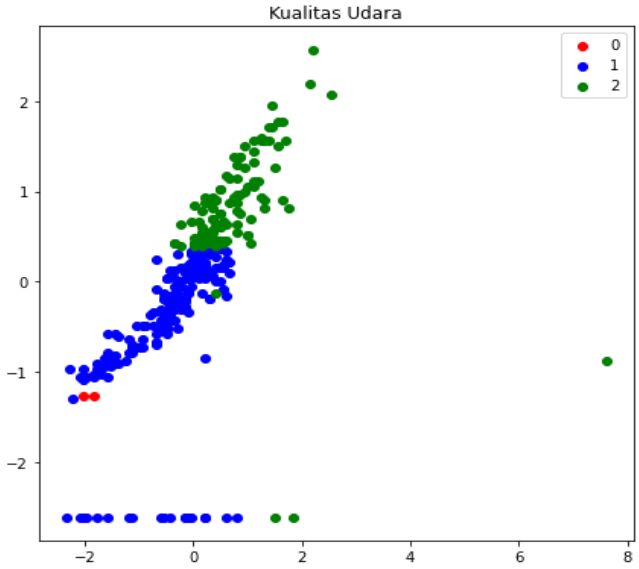

Figure 3. Classification results of dataset training

Figure 4 below shows the results of data classification based on air quality after testing data.

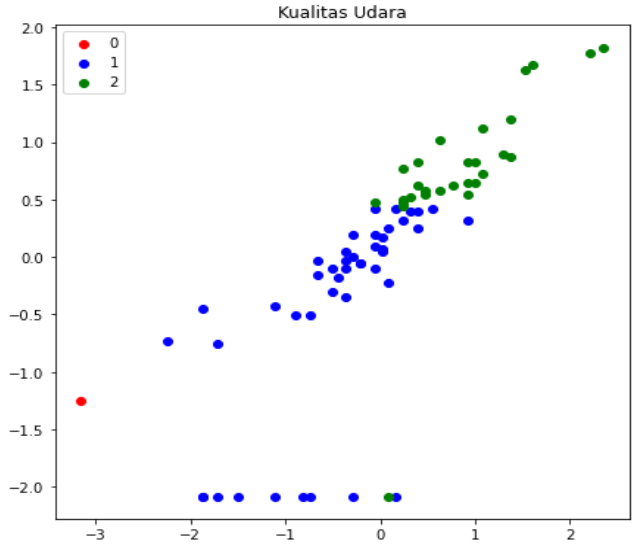

Figure 4. Classification results of dataset testing

At the partitioning stage (training and data testing) two treatments were given in the partitioning method, namely by using random sampling and stratified sampling. Then, the next step is to create an SVM classifier with two treatments in the kernel function, namely polynomial and

RBF. Table 4 below shows the results of the accuracy of the SVM method with some of these treatments.

Table 4. The level of accuracy in the SVM method using error analysis

Partition Method	Ratio	Polynomial		RBF	
		MAE	RMSE	MAE	RMSE
Random Sampling	80:20	0.19	0.44	0.09	0.31
	70:30	0.18	0.43	0.11	0.33
	60:40	0.14	0.37	0.09	0.29
Stratified Sampling	80:20	0.23	0.51	0.04	0.20
	70:30	0.21	0.49	0.05	0.23
	60:40	0.19	0.46	0.06	0.25

B. Autoregressive Integrated Moving Average (ARIMA) Experiment Results

Some of the attributes or variables in the dataset used in the study are not stationary, namely Max and Category. Then, differentiation is performed once on data that is still not stationary. Comparison of the state of the dataset before and after differentiation is shown in Figures 5 and 6 below.

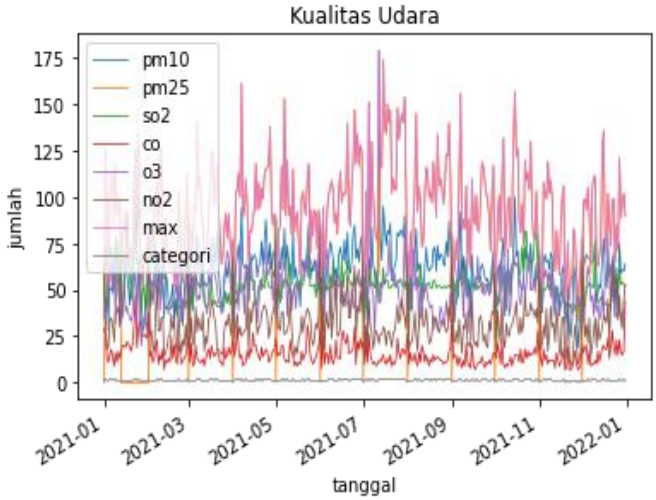

Figure 5. Dataset before differentiation (non-stationary)

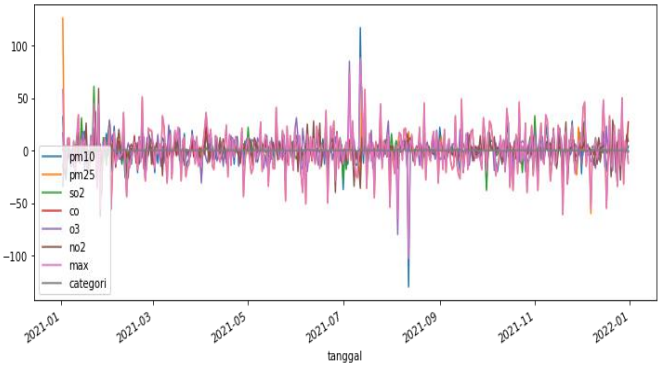

Figure 6. Dataset after differentiation (stationary)

After the data is stationary, ARIMA modeling is then carried out with parameters (1,1,1). The results of this parameter are obtained by using the auto_arima function in python. From the results of this modeling, forecasting is then carried out and the accuracy results are shown in Table 5.

Tabel 5. The level of accuracy in the ARIMA method using error analysis

Partition Method	Ratio	MAE	RMSE
Random Sampling	80:20	0.51	0.55
	70:30	0.58	0.61
	60:40	0.57	0.62

CONCLUSION

From the experiments result, several conclusions can be drawn:

1. In testing the dataset using SVM using random sampling and stratified sampling partitions produce different levels of accuracy as described above, that the random sampling partition with a ratio of 60:40 has the best accuracy, namely with RBF producing MAE 0.09 and RMSE 0.29 and with polynomial produces MAE 0.14 and RMSE.
2. In the SVM test, treatment using RBF produces a higher level of accuracy. This is because usually linear and polynomial kernels take less time and provide lower accuracy than rbf or Gaussian kernels.
3. In testing the dataset using ARIMA before differentiating several variables such as Max and Category it produces non-stationary data, but after differentiation with parameters (1,1,1) the dataset becomes stationary. The level of accuracy obtained in the use of random sampling partitions obtained a ratio of 80:20 which is the best, namely MAE of 0.51 and RMSE of 0.55.
4. From ARIMA and SVM testing, it can be said that SVM testing has higher accuracy results. This can be seen from the average accuracy results with several treatments.

From the testing, it is possible to do research again to test the dataset on ARIMA using a stratified sampling partition in the next test, so that the results of this study can be perfected.

REFERENCE

[1] S. Chapman, J. E. M. Watson, A. Salazar, M. Thatcher, and C. A. McAlpine, "The impact of urbanization and climate change on urban temperatures: a systematic review," *Landsc Ecol*, vol. 32, no. 10, pp. 1921–1935, Oct. 2017, doi: 10.1007/s10980-017-0561-4.

[2] L. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Front Public Health*, vol. 8, Feb. 2020, doi: 10.3389/fpubh.2020.00014.

[3] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep Distributed Fusion Network for Air Quality Prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2018, pp. 965–973. doi: 10.1145/3219819.3219822.

[4] A. M. Fiore *et al.*, "Global air quality and climate," *Chem Soc Rev*, vol. 41, no. 19, p. 6663, 2012, doi: 10.1039/c2cs35095e.

[5] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 2, no. 2, Aug. 2016, doi: 10.28932/jutisi.v2i2.476.

[6] W. Lu *et al.*, "Air pollutant parameter forecasting using support vector machines," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, IEEE, 2002, pp. 630–635. doi: 10.1109/IJCNN.2002.1005545.

[7] M. Dun, Z. Xu, Y. Chen, and L. Wu, "Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine," *Math Probl Eng*, vol. 2020, pp. 1–13, May 2020, doi: 10.1155/2020/8914501.

[8] A. Kaya, R. Ozturk, and C. Altin Gumussoy, "Usability Measurement of Mobile Applications with System Usability Scale (SUS)," 2019, pp. 389–400. doi: 10.1007/978-3-030-03317-0_32.

[9] Gourav, J. K. Rekhi, P. Nagrath, and R. Jain, "Forecasting Air Quality of Delhi Using ARIMA Model," 2020, pp. 315–325. doi: 10.1007/978-981-15-0372-6_25.

[10] Y. Zhang, H. Yang, H. Cui, and Q. Chen, "Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China," *Natural Resources Research*, vol. 29, no. 2, pp. 1447–1464, Apr. 2020, doi: 10.1007/s11053-019-09512-6.

[11] M. Awad and R. Khanna, *Efficient Learning Machines*. Berkeley, CA: Apress, 2015. doi: 10.1007/978-1-4302-5990-9.

[12] KNIME Community Hub, "KNIME Base nodes," *KNIME*. https://hub.knime.com/knime/extensions/org.knime.features.base/latest (accessed Jan. 01, 2023).