Mathematical Journal of Modelling and Forecasting
Vol. 3, No. 2, December 2025, pp. 36-41

Application of Algorithm Learning Vector Quantization For Air Quality Classification

Roufsaldiaz Nawfal[1] Dina Fitria[2,*], and Chairina Wirdiastuti[3]

[1,2,3] Department of Statistics, Universitas Negeri Padang, Indonesia
[*]dinafitria@fmipa.unp.ac.id

Abstract. This study aims to classify air quality using the Learning Vector Quantization (LVQ) algorithm based on the Air Quality and Pollution Assessment dataset obtained from Kaggle. The dataset comprises 5,000 observations, of which 4,000 were used for training and 1,000 for testing. The analytical process includes data preprocessing (normalization), the construction and training of the LVQ model, and performance evaluation using a confusion matrix. The experimental results demonstrate that the LVQ model successfully classified 903 of 1,000 test samples, yielding an overall accuracy of 90.3%. This level of accuracy indicates that the LVQ algorithm can capture relevant patterns in air quality variables and perform reliable classification across different air quality categories. The findings suggest that LVQ can serve as a potential foundation for developing automated air quality monitoring and decision-support systems. Future studies are encouraged to compare LVQ with other machine learning classification techniques to build a more optimal model and to gain deeper analytical insights.

Keywords: Air Quality, Classification, Learning Vector Quantization.

1 Introduction

Air is an invisible element that plays a vital role in life on Earth, because its oxygen content is essential for living things. Although air is invisible, its effects can be felt. Clean air is healthy for the respiratory tract, reduces the risk of long-term illness, helps prolong life, increases endurance and focus, and improves mood [1]. On the other hand, polluted air or air that falls below air quality standards can harm animals and plants, disrupt comfort, and threaten human health [2]. Therefore, understanding air quality and conditions is increasingly important for anticipating the adverse impacts they can cause.

Air quality is an essential indicator of environmental health and human welfare. The decline in air quality due to increased air pollution has become a global problem with severe impacts on public health and ecosystems. According to a report by the World Health Organization (WHO), approximately 7 million premature deaths each year are attributable to air pollution, particularly in urban areas with dense industrial and transportation activities [3]. This situation makes air pollution a serious threat that requires a deeper understanding, particularly with respect to the various pollutants that can cause health impacts.

Air pollution causes health problems for many people worldwide due to the compounds it contains. Compounds monitored in air quality, such as particulate matter (PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), are pollutants that can affect human health and the environment [4]. The effects of exposure to these pollutants are further explained in a study [5], which reports that, in the short term, air pollution can cause eczema, acne, premature aging, heart attacks, asthma, and ADHD in infants from the womb through childhood. In the long term, air pollution can cause blood vessel blockages, premature birth, pneumonia, lung cancer, asthma, Alzheimer's, Parkinson's, stroke, and cognitive decline.

The impact of air pollution can be minimized by providing an effective air quality monitoring and analysis system. In this process, a method is needed that can accurately classify air quality levels based on existing pollutant parameters. This approach can be done using a classification method. Classification is the process of grouping objects or data into specific categories or classes based on specific characteristics or attributes [6]. A relevant approach to environmental classification problems is an artificial intelligence-based learning method, such as artificial neural networks. Artificial neural networks are processing systems designed and trained to learn by adjusting their weights to solve complex problems [7]. In artificial neural networks, several

methods use supervised training, including Boltzmann, Hopfield, Backpropagation, and Learning Vector Quantization. Of these four methods, the only one with fast training and execution times is Learning Vector Quantization (LVQ), making LVQ the best among them [8]. LVQ works by comparing the distance between the input data and the feature vector (prototype) for each class, thereby grouping the data into the closest category.

Previous studies have shown that, in addition to fast training and execution times, LVQ can produce high model accuracy. As in study [9], the classification of river water quality achieved an accuracy of 81.13%. Furthermore, in study [10], in classifying air quality in the city of Pekanbaru, an accuracy of 95.23%. These findings indicate that LVQ is a suitable approach for classification problems, as its competitive learning characteristics enable the model to learn data patterns more effectively.

Therefore, this study will apply the LVQ algorithm to classify air quality using the Air Quality and Pollution Assessment dataset from Kaggle. This dataset contains information on temperature, humidity, concentrations of PM2.5, PM10, NO2, SO2, CO, distance to industrial areas, and population density. By applying the LVQ method, it is hoped that an accurate classification model will be obtained in determining air quality categories. In addition, this study aims to assess the effectiveness of LVQ for classifying air quality.

2 Research Methods

2.1 Data Source and Research Variable

The data used in this study is secondary data sourced from the website https://www.kaggle.com/. The dataset used is the "Air Quality and Pollution Assessment," which contains 5000 observations. This dataset includes 10 attributes with 1 "target" variable used to indicate whether the air quality in an area is good, moderate, unhealthy, or hazardous. The descriptions of the overall variables are presented in Table 1.

Table 1. Variable description

Variable	Description	Unit
X_1	Average temperature of the region	°C
X_2	Relative humidity recorded in the region	%
X_3	Fine particulate matter levels	µg/m³
X_4	Coarse particulate matter levels	µg/m³
X_5	Nitrogen dioxide levels	ppb
X_6	Sulfur dioxide levels	ppb
X_7	Carbon monoxide levels	ppm
X_8	Distance to the nearest industrial zone	km
X_9	Number of people per square km in the region	people/km²
Y	Good, Moderate, Poor, Hazardous	

2.2 Data Analysis Techniques

This study employs the LVQ algorithm in RStudio. The LVQ algorithm is an Artificial Neural Network algorithm whose output values are classified. The Learning Vector Quantization method is used for classification, where the number of classes is predetermined. The advantage of the LVQ method is its ability to train competitive layers, thereby enabling it to classify input vectors automatically. The steps in the LVQ algorithm are as follows [11]:

1. Data input
2. Data normalization, this step aims to change the scale of variable values in the range of 0 to 1 without changing their relative distribution. The data normalization calculation is formulated as follows:

$$X_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Description:

X_i' : normalized value of the *i*-th variable
x_i : actual value of the *i*-th variable
x_{min} : minimum value of the variable
x_{max} : maksimum value of the variable

3. Divide the training data into 80% and the testing data into 20%. Training data is a subset of the data used to learn patterns in input variables and their relationships with the target (class). Meanwhile, test data are used to evaluate the model's classification performance. In general, data are divided at a 80%:20% ratio.

4. Perform classification using the LVQ algorithm

   a. Initialize values:

      1) Initial weights (w_{ij}) where w_{ij} is the weight of the *i*-th neuron for the *j*-th input variable
      2) Maxepoch
      3) Learning Rate (α)
      4) Minimum expected error (Eps)

   b. Set the initial condition *epoch* = 0

   c. Perform the process if ($epoch < maksimum\ epoch$) and ($\alpha > eps$):
      1) *Epoch* = *epoch* + 1;
      2) Compute the Euclidean distance between the input vector X and each weight vector W_i then determine the minimum distance C_j, using the following formula:

$$d(X, W_i) = \sqrt{\sum_{j=1}^{n} (X_i - X_{ij})^2} \quad (2)$$

      Description:

      $d(X, W_i)$: minimum distance to the weight vector.
      X : input vector.
      W_i : weight vector of the *i*-th neuron in the output layer.
      j : index of the variable.

   d. Update the weights according to the following rules:
      1) If $T = C_j$ then:

$$W_j(new) = W_j(old) + \alpha(X_i - W_j(old)) \quad (3)$$

      2) If $T \neq C_j$ Then:

$$W_j(new) = W_j(old) - \alpha(X_i - W_j(old)) \quad (4)$$

      Description:

      X : training vector.
      T : actual class of the training vector.
      W_j : weight vector of the *j*-th output neuron.
      C_j : class represented by the *j*-th output neuron.

5. Determine the accuracy level of the LVQ method using a confusion matrix.

   A confusion matrix is used to measure and evaluate the performance of a classification model [12]. The confusion matrix is shown in Table 2.

Table 2. Confusion matrix table

Classification		Prediction Class			
		Class = 0	Class = 1	Class = 2	Class = 3
Original Class	Class = 0	TP_0	FP_{01}	FP_{02}	FP_{03}
	Class = 1	FN_{10}	TP_1	FP_{12}	FP_{13}
	Class = 2	FN_{20}	FN_{21}	TP_2	FP_{24}
	Class = 3	FN_{30}	FN_{31}	FN_{32}	TP_3

Description:
True Positive (TP) : True original class with positive predicted class result
True Negative (TN) : True original class with negative predicted class result.
False Negative (FN) : False original class with negative predicted class result.
False Positive (FP) : False original class with positive predicted class result

The value calculated in the confusion matrix in this article is accuracy. Accuracy is used to evaluate the classification performance results in the formula presented in equation:

$$Accuration = \frac{Number\ of\ correct\ prediction}{Number\ of\ data\ test} \times 100\% \quad (5)$$

3 Results and Discussion

3.1 Data Normalization

Data normalization aims to minimize computational errors by scaling the data to the 0-1 range. The results of data normalization are shown in Table 3.

Table 3. Data normalization result

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	1	0.303	0.033	0.021	0.021	0.106	0.225	0.162	0.369	0.472
2	2	0.362	0.251	0.017	0.057	0.200	0.301	0.348	0.163	0.170
:	:	:	:	:	:	:	:	:	:	:
5000	2	0.236	0.454	0.276	0.299	0.274	0.326	0.237	0.248	0.355

Based on the normalization results in Table 3, all pollutant variables and environmental characteristics have been successfully transformed into a range of values from 0 to 1 without changing the distribution pattern of the original data. From the resulting value patterns, it is evident that most pollutant variable values fall within the low range, whereas only a small number of observations have high values near 1. This condition indicates a right-skewed distribution (positive skewness), a common characteristic of environmental pollution data, in which incidents with high pollution levels are relatively rare compared with normal conditions.

Although the normalization process does not eliminate distributional dilution, it ensures equal scaling across variables, preventing any single variable from dominating the distance calculation in the Learning Vector Quantization (LVQ) algorithm. This is important because LVQ uses Euclidean distance as its learning criterion. Unlike standardization, which adjusts data by subtracting the mean and dividing by the standard deviation and is more suitable for near-normal distributions, normalization is used in this study because it is more stable across varying value ranges and the presence of extreme values, thereby supporting a more consistent classification process.

3.2 Training Data and Testing Data

The data consist of daily checks of whether the air quality in a region falls into the categories of Good, Moderate, Poor, or Hazardous, totaling 5000 samples. The data were then split into 80% of the sample (4000 samples) as the training set and 20% (1000 samples) as the test set. The results of the data division are shown in Table 4.

Table 4. Splitting Data

Data Group	Number of Samples	Percentage
Training	4000	80%
Testing	1000	20%

3.3 Learning Vector Quantization

The LVQ algorithm is a learning algorithm that offers greater flexibility and faster classification of multi-class data with both numerical and categorical features. It has no data limitations and provides high classification accuracy [1]. In this study, LVQ is applied to minimize classification errors for air quality into four categories: Good, Moderate, Unhealthy, and Hazardous. The classification results are presented in Table 5.

Table 5. Confusion Matrix Result

Classification		Predicted Class			
		Good	Moderate	Poor	Hazardous
Original Class	Good	388	9	0	0
	Moderate	3	281	26	0
	Poor	0	20	152	24
	Hazardous	0	0	15	82

The results of testing 1,000 test data are shown in the confusion matrix in Table 5. The LVQ model correctly classified 903 data points, yielding an accuracy of 90.3%. This value indicates that LVQ can effectively recognize air quality patterns in multivariate numerical data.

When evaluated per class, the model performs best on the Good class, where most data are predicted accurately. This shows that air conditions with low pollution levels have consistent characteristics. In the Moderate and Poor classes, several classification errors occur across both classes, indicating similar characteristics in pollutant variable values and making the class boundaries less clear. In the Hazardous class, a small portion of the data was predicted as Poor, indicating that some extreme air conditions exhibit characteristics similar to that category. Overall, the model demonstrates a strong ability to recognize hazardous air conditions.

The results of this study are in line with previous studies that applied the LVQ algorithm to environmental quality classification problems. As in study [2], an accuracy of 81.13% was achieved in river water quality classification, whereas in study [5], an accuracy of 95.23% was achieved in air quality classification in the city of Pekanbaru. In this study, an accuracy of 90.3% was achieved, indicating that LVQ exhibits competitive and stable performance, even when applied to datasets with varying numbers of classes and variables. Differences in accuracy can be influenced by complexity, the number of classes, and the characteristics of the area under analysis.

In this study, accuracy is used as the primary evaluation metric because the dataset's class distribution is relatively balanced. Under balanced data conditions, accuracy can fairly and comprehensively represent the model's performance across all classes. Other evaluation metrics, such as precision, recall, and F1-score, are generally more appropriate when class imbalance is present or when the analysis focuses on the performance of a particular class. Therefore, accuracy is considered the most suitable metric for this study, which aims to evaluate the LVQ model's overall performance in classifying air quality.

4 Conclusion

The results of this study show that the LVQ algorithm can classify air quality with good performance. Of the 5000 available samples, 4000 were used for training and 1000 for testing. During the testing phase, the model correctly classified 903 of 1000 data points, yielding an accuracy of 90.3%. This achievement indicates that the LVQ model can effectively recognize patterns in air quality data and has the potential to serve as a basis for developing air quality classification systems across various regions. For further research, it is recommended to explore alternative classification methods, such as Categorical Boosting or other modern machine learning algorithms, to identify an approach with higher accuracy in classifying air quality.

Reference

[1] A. A. H. Kirono, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma Naive Bayes," eProceedings of Engineering, vol. 9, no. 3, Jun. 2022, Accessed: Jan. 05, 2026. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/18002

[2] R. F. Ramadhani, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN," Journal of Computer System and Informatics (JoSYC), vol. 3, no. 4, pp. 536–543, Sep. 2022, doi: 10.47065/JOSYC.V3I4.2117.

[3] World Health Organization, "Ambient (outdoor) air pollution," World Health Organization. Accessed: Jan. 05, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] A. D. Wiranata, S. Soleman, I. Irwansyah, I. K. Sudaryana, and R. Rizal, "Klasifikasi Data Mining untuk Menentukan Kualitas Udara di Provinsi DKI Jakarta Menggunakan Algoritma K-Nearest Neighbors (K-NN)," Infotech: Journal of Technology Information, vol. 9, no. 1, pp. 95–100, Jun. 2023, doi: 10.37365/JTI.V9I1.164.

[5] N. F. Khusna, S. Aulia, S. Amaria, A. Rahmah, S. A. Sanmas, and F. Fauzi, "Peramalan Kualitas Udara di Semarang Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA)," Prosiding Seminar Nasional Unimus, vol. 6, no. 0, Nov. 2023, Accessed: Jan. 05, 2026. [Online]. Available: https://prosiding.unimus.ac.id/index.php/semnas/article/view/1484

[6] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann Publishers, 2011. doi: 10.1016/C2009-0-61819-5.

[7] E. Setyowati and S. Mariani, "Penerapan Jaringan Syaraf Tiruan dengan Metode Learning Vector Quantization (LVQ) untuk Klasifikasi Penyakit Infeksi Saluran Pernapasan Akut (ISPA)," PRISMA, Prosiding Seminar Nasional Matematika, vol. 4, pp. 514–523, Feb. 2021, Accessed: Jan. 05, 2026. [Online]. Available: https://journal.unnes.ac.id/sju/prisma/article/view/44356

[8] S. Ding, X. H. Chang, and Q. H. Wu, "A study on the application of learning vector quantization neural network in pattern classification," Applied Mechanics and Materials, vol. 525, pp. 657–660, 2014, doi: 10.4028/WWW.SCIENTIFIC.NET/AMM.525.657.

[9] R. Hamidi, M. T. Furqon, and B. Rahayudi, "Implementasi Learning Vector Quantization (LVQ) untuk Klasifikasi Kualitas Air Sungai," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 1, no. 12, pp. 1758–1763, Aug. 2017, Accessed: Jan. 05, 2026. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/635

[10] J. Hendri. Klasifikasi Kualitas Udara Menggunakan Algoritma Learning Vector Quantization di Kota Pekanbaru. Skripsi thesis, Universitas Islam Negeri Sultan Syarif Kasim Riau. 2021

[11] H. Harliana and S. Kirono, "Penerapan Learning Vector Quantization Dalam Memprediksi Jumlah Rumah Tangga Miskin," Jurnal Sains dan Informatika, vol. 5, no. 2, pp. 118–127, Dec. 2019, doi: 10.34128/JSI.V5I2.192.

[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf Process Manag, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.