Klasifikasi Indeks Standar Pencemaran Udara untuk Data Tidak Seimbang menggunakan Pendekatan Pembelajaran Mesin

Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach

[1]Bryan Valentino Jayadi, [2]Manatap Dolok Lauro, [3]Zyad Rusdi, [4]Teny Handhayani*
[1,2,3,4]Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta, Indonesia
*e-mail: tenyh@fti.untar.ac.id

(***received*****:** 10 September 2023, ***revised*****:** 16 Januari 2024, ***accepted*****:** 16 Januari 2024)

Abstrak

Pencemaran udara merupakan salah satu masalah yang ada di masyarakat. Polusi udara mempengaruhi kesehatan manusia dan lingkungan. Di Indonesia, indeks standar pencemar udara (ISPU) diukur dari kadar partikulat 10 (PM_{10}), karbon monoksida (CO), sulfur dioksida (SO_2), ozon (O_3), dan nitrogen dioksida (NO_2). Penelitian ini dilakukan untuk mengevaluasi kinerja algoritma pembelajaran mesin, misalnya, Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree, dan AdaBoost, untuk mengklasifikasikan indeks kualitas udara berdasarkan nilai PM_{10}, CO, SO_2, O_3, dan NO_2 dengan jumlah sampel yang tidak seimbang. Kualitas udara diklasifikasikan menjadi Baik, Sedang, dan Tidak Sehat. Dataset diunduh dari Open Data Jakarta dari tahun 2010 - 2021. Data yang berisi 4383 sampel terdiri dari 1155 sampel Baik, 3087 sampel Sedang, dan 141 sampel Tidak Sehat. Hasil eksperimen menunjukkan bahwa Decision Tree mengungguli metode lainnya. Decision Tree menghasilkan akurasi, presisi, recall, dan skor F1 masing-masing sebesar 99%, 98%, 99%, dan 98%.

Kata kunci: ISPU; klasifikasi; data tidak seimbang; pembelajaran.

Abstract

Air pollution is one of the problems in society. Air pollutions affect human health and environment. In Indonesia, air quality index is measured by the level of particulate matter 10 (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and nitrogen dioxide (NO_2). This research is conducted to evaluate the performance of machine learning algorithms, e.g., Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree, and AdaBoost, to classify air quality index based on the level of PM_{10}, CO, SO_2, O_3, and NO_2 with imbalanced samples. The air quality index is classified into Good, Moderate, and Unhealthy. The dataset is downloaded from Open Data Jakarta from 2010 -2021. The data containing 4383 samples consist of 1155 samples of Good, 3087 samples of Moderate, and 141 samples of Unhealthy. The experimental results show that Decision Tree outperforms other methods. Decision Tree produces accuracy, precision, recall, and F1-score of 99%, 98%, 99%, and 98%, respectively.

Keywords: *air quality index; classification; imbalanced data, machine learning.*

1 Introduction

Air quality is one of the important issues in society that needs properly managed. Air Quality Index is measured by the level of carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozon (O_3), particulate matter 10 (PM_{10}), and particulate matter 2.5 ($PM_{2.5}$). Air pollution has negative impact on human health and environment. A study said that air pollution affects lung and heart disease that possibly leads to premature death [1]. Particulate matter possibly causes some disease, e.g., asthma and lung disease [2]. A study about low- and middle-income countries found that

 ISSN:2302-8149
e-ISSN:2540-9719

industrial development and urbanization in a very short period might cause air pollution [3]. A research reveal that industrial air pollution indirectly causes crop yield decrease [4]. Air pollution also affects cattle mortality during summer season [5]. A study found that $PM_{2.5}$ associated with cow's milk production [6].

Air pollution in Indonesia becomes one of important issues for public health. The Ministry of Environment and Forestry in the Republic of Indonesia classifies the air quality index as good (1-50), moderate (51-100), unhealthy (101-200), very unhealthy (201-300), and hazardous (over 300) [7]. Traffic emissions, dust emission, peat fire, and forest fire are the majority factors that contribute to the air pollution in Indonesia [8]. A study about air pollution in Jakarta found that the high level of O_3, $PM_{2.5}$, and PM_{10} contributes worse pollution in Jakarta [9]. Worst air pollution in Jakarta possible causes respiratory disease, mortality, and adverse health impact on children [10]. A policy of large-scale social restriction in Jakarta during Covid-19 outbreak successfully decrease the air pollution index of some pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, O_3), and there was a positive correlation of SO_2, CO and $PM_{2.5}$ to Covid 19-cases [11].

Machine learning methods has been implemented to find solutions in air pollution issues [12][13]. Seasonal Auto Regressive Integrated Moving Average (SARIMA), Support Vector Regression, Long Short-Term Memory (LSTM), gated recurrent unit (GRU), random forest regression, linear regression are popular algorithms for forecasting the air pollutant values [9] [14], [15] [16] [17]. Some algorithms (Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, Artificial Neural Network (ANN)) have been implemented for classification jobs in air pollution research [18] [19].

This paper investigates the performance of some machine learning algorithms to classify the imbalanced data of air quality index. The research goal is to evaluate some machine learning algorithms to classify the data of air quality index when the samples are imbalanced. Imbalanced data means that the proportion of samples of the classes are skewed. The best model is then used to build an application to classify air quality index based on the level of PM_{10}, CO, SO_2, O_3, and NO_2.

2 Literature Review

Some popular classification algorithms are AdaBoost, Naïve Bayes, Logistic Regression, Decision Tree, and Support Vector Machine. The performance of classification algorithms can be measured using precision, recall, accuracy, and F1 score.

AdaBoost

Boosting algorithm was a powerful tool for binary class classifiers developed by Freund and Schapire [20]. Multi-class AdaBoost was introduced for multi-class classification [21]. Let $(x_1, c_1), \dots, (x_n, c_n)$ be a pair of input variables and target. The class target c_i has a finite value of $\{1,2, \dots, K\}$. The goal of the algorithm is to find the class c from $\{1,2, \dots, K\}$ for input $\boldsymbol{x}$. The mathematical models of AdaBoost can be described using equations (1), (2), (3), and (4).

$$err^{(m)} = \sum_{i=1}^{n} w_i \mathbb{I}\left(c_i \neq T^{(m)}(x_i)\right) \Big/ \sum_{i=1}^{n} w_i \quad (1)$$

$$\alpha^{(m)} = \log\frac{1 - err^m}{err^m} + \log(K-1) \quad (2)$$

$$w_i \leftarrow w_i . exp\left(\alpha^{(m)} . \mathbb{I}\left(c_i \neq T^{(m)}(x_i)\right)\right), \text{for } i = 1, 2, \dots, n \quad (3)$$

$$C(x) = \arg\max_k \sum_{m=1}^{M} \alpha^{(m)} . \mathbb{I}\left(T^{(m)}(x) = k\right) \quad (4)$$

Naïve Bayes

Naive Bayes is a classifier that applies maximum likelihood. Let $X = (A_1, A_2, \dots, A_k)$ be a sample vector and A_j be the jth variable that has some values x_j. Naive Bayes classifier can be defined by equations (5), (6), (7), and (8), where D is sample set, $N(D)$ is the total number of samples, and $N(C_i)$ is the number of samples in class C_i [22].

 ISSN:2302-8149
e-ISSN:2540-9719

$$P(X|C_i) = \prod_{j=1}^{k} P(A_j = x_j|C_i) \quad (5)$$

$$P(C_i|X) = \frac{\prod_{j=1}^{k} P(A_j = x_j|C_i)P(C_i)}{P(X)} \quad (6)$$

$$P(C_i|X) = \alpha \prod_{j=1}^{k} P(A_j = x_j|C_i)P(C_i), \text{ where } \frac{1}{P(X)} = \alpha(> 0) \quad (7)$$

$$P(C_i|X) = \alpha \prod_{j=1}^{k} \frac{N(C = C_i, A_j = x_j)}{N(C_i)} . \frac{N(C_i)}{N(D)} \quad (8)$$

Logistic Regression

Logistic regression is a function that map features to the target to predict the probability when a new item belongs to one of the classes [23]. Original logistic regression is a method for binary classifier [24]. Logistic regression for multiclass classification is defined by equation (9), where $a_k = w_k^T \phi$ and $P(C_k|\phi)$ is probability of class C_k[25]. The likelihood function for multiclass logistic regression can be computed using equations (9) and (10).

$$P(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad (9)$$

$$P(T|w_1, \dots, w_k) = \prod_{n=1}^{N} \prod_{k=1}^{K} P(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}} \quad (10)$$

Decision Tree

A decision tree is an algorithm works for classification task [26]. A decision tree contains nodes that establish a directed tree. Two main phases in the decision tree algorithm are developing a tree as a model and classification. The first step is creating an empty tree and then creating each node from a selected variable using a particular measurement. The process is finished when the tree reaches the leaves. The classification process is started from the root of the tree and follows the path suitable to the values of observed variables until reaches the leaf. The information gain used for attribute selection and is defined using equations (11), (12), and (13), where T is training set and A is attribute. The detail implementation of the decision tree can be found in the previous research [27].

$$Gain(T, A_k) = E(T) - E_{A_k}(T) \quad (11)$$

$$E(T) = -\sum_{i=1}^{n} \frac{n(C_i, T)}{|T|} \log_2 \frac{n(C_i, T)}{|T|} \quad (12)$$

$$E_{A_k}(T) = \sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} E(T_v^{A_k}) \quad (13)$$

Support Vector Machine

Support Vector Machine (SVM) is one of algorithms for classification task [28]. The basic idea in SVM is to separate two clusters with minimum error using hyperplane. The hyperplane is defined by $(\omega \cdot x) + b = 0$. There are two problems in SVM: linearly separable and non-linear separable problems. The optimal linear separation can be computed using equation $\phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w)$. In non-linear separable problem, SVM maps the input vector to a high-dimensional feature space and arrange optimal separator in the feature space. Mapping input vectors to a high-dimensional can be done through kernel functions. Some popular kernel functions are linear kernel, RBF kernel, and Polynomial kernel. Linear, RBF, and Polynomial kernel can be defined using equations (14), (15), and

 ISSN:2302-8149
e-ISSN:2540-9719

(16), where $X = (x_1, x_2, \dots, x_k)$ is an input vector [29]. SVM for multi class classification applies one-against-rest or one-against-one method.

$$K(x_i, x_j) = x_i \cdot x_j^T \quad (14)$$

$$K(x_i, x_j) = exp\left(-exp\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \text{ where } \sigma > 0 \quad (15)$$

$$K(x_i, x_j) = \left(x_i^T x_j + c\right)^d \quad (16)$$

Evaluation Metrics

Evaluation metrics for classification are precision, recall, accuracy, F1-Score, and confusion matrix. Suppose, TP, TN, FP, FN, N, and m denote true positive, true negative, false positive, false negative, the number of samples, and the number of classes. Precision, recall, accuracy, F1-Score can be computed using equations (17, (18), (19), and (20) [30]. A confusion matrix describes a recap of the performance of classification algorithms related to the test data [31]. A confusion matrix is a matrix containing the number of true objects correctly classified and the number of objects that are misclassified.

$$precision = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FP_i} \quad (17)$$

$$recall = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FN_i} \quad (18)$$

$$accuracy = \frac{\sum_{i=1}^{m} TP_i + \sum_{i=1}^{m} TN_i}{N} \quad (19)$$

$$F1\ score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \quad (20)$$

3 Research Method

A research workflow is described in Figure 1. The first step is collecting the dataset from Open Data Jakarta. The second step is data preparation. This step contains missing values checking, data synchronization, and distributing data for training and testing. The third step is models training. It runs some machine learning algorithms, e.g., SVM, Logistic Regression, Decision Tree, Naïve Bayes, and Ada Boost. The trained models are evaluated using testing data and their performances are measured using evaluation metrics for classifications (accuracy, precision, recall, F1- scores, confusion matrix). The trained models are then used to develop an application for AQI classification.

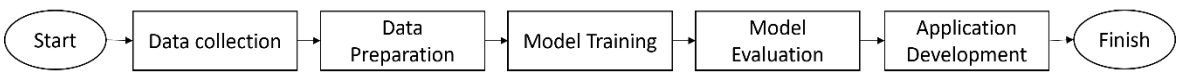

Figure 1 Research Workflow.

Results and Analysis

This paper uses dataset of air quality index from Open Data Jakarta from 2010 – 2021 contaning 4383 samples. The distribution of samples for each class are 1155 samples of good, 3087 samples of moderate, and 141samples of unhealthy. There are no missing values in the dataset. The distribution of samples across the classes is imbalanced. The majority samples are Moderate of 70.4%, followed by Good of 26.4%, and the least of all is Unhealthy of 3.2%. The data is then divided into 80% training data and 20% testing data. This paper runs experiment using SVM, Logistic Regression, Decision Tree, Naïve Bayes, and Ada Boost. It applies scikit-learn library for python [32]. The performance of algorithms is measured using accuracy, precision, recall, and F1-score.

 ISSN:2302-8149
e-ISSN:2540-9719

Table 1 Evaluation Metrics

No	Algorithm	Accuracy	Precision	Recall	F1-Score	Note
1	SVM	97%	97%	96%	97%	Polynomial C = 1.0
2	SVM	97%	96%	97%	97%	Polynomial C = 10
3	SVM	97%	96%	97%	96%	Polynomial C = 100
4	SVM	96%	98%	94%	96%	RBF = 1.0
5	SVM	97%	96%	95%	96%	RBF C = 10
6	SVM	98%	97%	97%	97%	RBF C = 100
7	SVM	92%	93%	89%	91%	Linear C = 1.0
8	SVM	92%	93%	89%	91%	Linear C = 10
9	SVM	92%	93%	89%	91%	Linear C = 100
10	Naïve Bayes	92%	90%	87%	88%	
11	Logistic Regression	88%	90%	87%	88%	
12	Decision Tree	99%	98%	99%	98%	
13	Ada Boost	82%	86%	91%	87%	

The experiments run several algorithms, e.g., SVM, Naïve Bayes, Logistic Regression, Decision Tree, and AdaBoost. The experiments using SVM apply linear kernel, RBF kernel, and Polynomial kernel with parameters $C = \{1, 10. 100\}$ and one-against-one model. Table 1 shows the evaluation metrics of SVM, Naïve Bayes, Logistic Regression, Decision Tree and AdaBoost. All algorithms obtain accuracy, precision, recall, and F1-score more than 82% Decision tree outperforms other methods. Decision tree shows slightly better performance than SVM using RBF kernel with C = 100. Various SVMs obtain accuracy, precision, recall, and F1-score over 89%. SVM applies RBF and Polynomial kernel has a higher evaluation score than SVM with linear kernel. Compared to other methods, AdaBoost produces the lower evaluation scores.

Table 2 Confusion matrix produced by Support Vector Machine

	Good	Moderate	Unhealthy	Total
Good	217	11	0	228
Moderate	5	609	2	616
Unhealthy	0	1	32	33
Total	227	620	30	877

Table 2 shows a confusion matrix produced by SVM using RBF kernel with C = 100. The highest miss-classification happened when 11 samples Good are identified as Moderate. Five samples of Moderate are misclassified as Good and two samples wrongly identified as Unhealthy. One sample of Unhealthy is miss-identified as Moderate.

Table 3 Confusion matrix produced by Naïve Bayes

	Good	Moderate	Unhealthy	Total
Good	196	32	0	228
Moderate	24	588	4	616
Unhealthy	0	7	26	33
Total	220	627	30	877

Table 3 displays the confusion matrix for Naïve Bayes. The highest miss-identified happened when 32 samples of Good class are misclassified as Moderate. The samples of Moderate have been misidentified to be Good as 24 samples and 4 be Unhealthy. Seven samples from Unhealthy are miss-identified as Moderate.

Table 4 Confusion matrix produced by Logistic Regression

 ISSN:2302-8149
e-ISSN:2540-9719

	Good	Moderate	Unhealthy	Total
Good	191	37	0	228
Moderate	26	569	21	616
Unhealthy	0	22	11	33
Total	217	628	32	877

A confusion matrix for Logistic Regression is displayed in Table 4. The highest miss-identified occurred when 37 samples Good are misidentified as Moderate. The samples of Moderate are misclassified as Good and Unhealthy at 26 and 21, respectively. A number of 22 samples of Unhealthy are misidentified as Moderate.

Table 5 Confusion matrix produced by AdaBoost

	Good	Moderate	Unhealthy	Total
Good	221	7	0	228
Moderate	149	467	0	616
Unhealthy	0	0	33	33
Total	221	7	0	228

Table 5 shows confusion matrix for AdaBoost. The highest misclassification happened when 149 samples of Moderate were identified as Good. Seven samples of Good are misclassified as Moderate.

Table 6 Confusion Matrix produced by Decision Tree

	Good	Moderate	Unhealthy	Total
Good	224	4	0	228
Moderate	6	609	1	616
Unhealthy	0	0	33	33
Total	230	613	34	

Table 6 displays confusion matrix of Decision Tree. Four samples of Good are misidentified as Moderate. The samples of Moderate are misclassified as Good and Unhealthy of 6 and 1, respectively.

Overall, AdaBoost classifier obtained the highest misclassified when 149 samples of Moderate are identified as Good. AdaBoost, Naive Bayes, and Logistic Regression have the higher misclassified in the samples of Moderate. Decision Tree performs the lowest misclassification. Decision Tree model is then used as a model to build an application to classify the air quality index. Figure 2 shows the application to classify the air quality index.

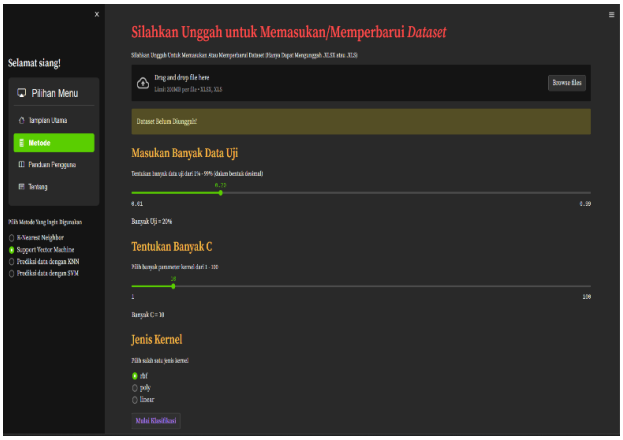

Figure 2 Application to classify air pollution index.

 ISSN:2302-8149
e-ISSN:2540-9719

4 Conclusion

In conclusion, Decision Tree model works well to classify the air quality index in imbalanced data. Decision Tree outperforms Support Vector Machine, Logistic Regression, Naïve Bayes, and AdaBoost to classify the air quality index based on the values of PM_{10}, CO, SO_2, O_3, and NO_2. The Decision Tree produces accuracy, precision, recall, and F1-score of 99%, 98%, 99%, and 98%, respectively. Future work is to analyse the factors contributes to the air quality index using causal learning approach.

Acknowledgement

This paper is supported by a research grant for undergraduate students from LPPM Universitas Tarumanagara.

Reference

[1] B. Ritz, B. Hoffmann, and A. Peters, "The Effects of Fine Dust, Ozone, and Nitrogen Dioxide on Health," *Dtsch Arztebl Int*, Dec. 2019, doi: 10.3238/arztebl.2019.0881.

[2] H. Chen *et al.*, "Effects of air pollution on human health–Mechanistic evidence suggested by in vitro and in vivo modelling," *Environ Res*, vol. 212, p. 113378, Sep. 2022, doi: 10.1016/j.envres.2022.113378.

[3] P. Mannucci and M. Franchini, "Health Effects of Ambient Air Pollution in Developing Countries," *Int J Environ Res Public Health*, vol. 14, no. 9, p. 1048, Sep. 2017, doi: 10.3390/ijerph14091048.

[4] W. Wei and Z. Wang, "Impact of Industrial Air Pollution on Agricultural Production," *Atmosphere (Basel)*, vol. 12, no. 5, p. 639, May 2021, doi: 10.3390/atmos12050639.

[5] B. Cox, A. Gasparrini, B. Catry, F. Fierens, J. Vangronsveld, and T. Nawrot, "Cattle mortality as a sentinel for the effects of ambient air pollution on human health," *Archives of Public Health*, vol. 73, no. S1, p. P22, Dec. 2015, doi: 10.1186/2049-3258-73-S1-P22.

[6] B. L. Beaupied *et al.*, "Cows as canaries: The effects of ambient air pollution exposure on milk production and somatic cell count in dairy cows," *Environ Res*, vol. 207, p. 112197, May 2022, doi: 10.1016/j.envres.2021.112197.

[7] KLHK, "Indeks Standar Pencemar Udara (ISPU) Sebagai Informasi Mutu Udara Ambien di Indonesia."

[8] N. A. Istiqomah and N. N. N. Marleni, "Particulate Air Pollution in Indonesia: quality index, characteristic, and source identification," *IOP Conf Ser Earth Environ Sci*, vol. 599, no. 1, p. 012084, Nov. 2020, doi: 10.1088/1755-1315/599/1/012084.

[9] T. Handhayani, "An integrated Analysis of Air Pollution and Meteorological Conditions in Jakarta," *Sci Rep*, vol. 13, no. 1, p. 5798, Apr. 2023, doi: 10.1038/s41598-023-32817-9.

[10] G. Syuhada *et al.*, "Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia," *Int J Environ Res Public Health*, vol. 20, no. 4, p. 2916, Feb. 2023, doi: 10.3390/ijerph20042916.

[11] M. Rendana and L. N. Komariah, "The Relationship between Air Pollutants and COVID-19 Cases and its Implications for Air Quality in Jakarta, Indonesia," *Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan (Journal of Natural Resources and Environmental Management)*, vol. 11, no. 1, pp. 93–100, Apr. 2021, doi: 10.29244/jpsl.11.1.93-100.

[12] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine Learning-based Prediction of Air Quality," *Applied Sciences*, vol. 10, no. 24, pp. 1–17, Dec. 2020, doi: 10.3390/app10249151.

[13] M. Méndez, M. G. Merayo, and M. Núñez, "Machine Learning Algorithms to Forecast Air Quality: a survey," *Artif Intell Rev*, vol. 56, no. 9, pp. 10031–10066, Sep. 2023, doi: 10.1007/s10462-023-10424-4.

[14] T. Handhayani, I. Lewenusa, D. E. Herwindiati, and J. Hendryli, "A Comparison of LSTM and BiLSTM for Forecasting the Air Pollution Index and Meteorological Conditions in Jakarta," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2022, pp. 334–339. doi: 10.1109/ISRITI56927.2022.10053078.

[15] N. N. Maltare and S. Vahora, "Air Quality Index PrCediction using Machine Learning for Ahmedabad ity," *Digital Chemical Engineering*, vol. 7, pp. 1–9, Jun. 2023, doi: 10.1016/j.dche.2023.100093.
[16] K. Saikiran, G. Lithesh, B. Srinivas, and S. Ashok, "Prediction of Air Quality Index using Supervised Machine Learning Algorithms," in *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, IEEE, Sep. 2021, pp. 1–4. doi: 10.1109/ACCESS51619.2021.9563323.
[17] Z. Zhao, J. Wu, F. Cai, S. Zhang, and Y.-G. Wang, "A hybrid Deep Learning Framework for Air Quality Prediction with Spatial Autocorrelation during the COVID-19 Pandemic," *Sci Rep*, vol. 13, no. 1, pp. 1–17, Jan. 2023, doi: 10.1038/s41598-023-28287-8.
[18] I. I. Ridho and G. Mahalisa, "Analisis Klasifikasi Dataset Indeks Standar Pencemaran Udara (ISPU) di Masa Pandemi menggunakan Algoritma Support Vector Machine (SVM)," *Technologia : Jurnal Ilmiah*, vol. 14, no. 1, pp. 38–41, Jan. 2023, doi: 10.31602/tji.v14i1.8005.
[19] S. Syihabuddin Azmil Umri, "Analisis dan Komparasi Algoritma Klasifikasi dalam Indeks Pencemaran Udara di DKI Jakarta," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 2, pp. 98–104, Aug. 2021, doi: 10.33387/jiko.v4i2.2871.
[20] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of on-line Learning and an Application to Boosting," *J Comput Syst Sci*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.
[21] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Stat Interface*, vol. 2, pp. 349–360, 2009.
[22] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved Naive Bayes Classification Algorithm for Traffic Risk Management," *EURASIP J Adv Signal Process*, vol. 2021, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/s13634-021-00742-6.
[23] E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA: Apress, 2019. doi: 10.1007/978-1-4842-4470-8.
[24] E. Makalic and D. F. Schmidt, "Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems," in *Advances in Artificial Intelligence*, AI 2010., Berlin: Springer, 2010, pp. 213–222. doi: 10.1007/978-3-642-17432-2_22.
[25] W. H. Nugroho, S. Handoyo, Y. J. Akri, and A. D. Sulistyono, "Building Multiclass Classification Model of Logistic Regression and Decision Tree using the Chi-Square Test for Variable Selection Method," *Journal of Hunan University Natural Sciences*, vol. 49, no. 4, pp. 172–181, Apr. 2022, doi: 10.55463/issn.1674-2974.49.4.17.
[26] L. Rokach and O. Maimon, "Decision Trees," in *Data Mining and Knowledge Discovery Handbook*, New York: Springer-Verlag, 2005, pp. 165–192. doi: 10.1007/0-387-25465-X_9.
[27] I. Jenhani, N. Ben Amor, and Z. Elouedi, "Decision Trees as Possibilistic Classifiers," *International Journal of Approximate Reasoning*, vol. 48, no. 3, pp. 784–807, Aug. 2008, doi: 10.1016/j.ijar.2007.12.002.
[28] Y. Zhang, "Support Vector Machine Classification Algorithm and Its Application," 2012, pp. 179–186. doi: 10.1007/978-3-642-34041-3_27.
[29] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511809682.
[30] T. Handhayani, A. H. Pawening, and J. Hendryli, "An Automatic Recognition System for Digital Collections of Indonesian Traditional Houses using Convolutional Neural Networks for Cultural Heritage Preservation," *Int J Comput Intell Appl*, vol. 22, no. 02, Jun. 2023, doi: 10.1142/S1469026823500037.
[31] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2011, pp. 209–209. doi: 10.1007/978-0-387-30164-8_157.
[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2826–2830, 2011.