Original Article

Available online at
www.heca-analitika.com/ljes

Leuser Journal of Environmental Studies

Vol. 1, No. 2, 2023

Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring

Ghazi Mauer Idroes [1,2], Teuku Rizky Noviandy [3,*], Aga Maulana [3], Zahriah Zahriah [4], Suhendrayatna Suhendrayatna [5], Eko Suhartono [6], Khairan Khairan [7], Fitranto Kusumo [8], Zuchra Helwani [9] and Sunarti Abd Rahman [10]

[1] Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I)
[2] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia;
[3] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; trizkynoviandy@gmail.com (T.R.N.); agamaulana@usk.ac.id (A.M.)
[4] Department of Architecture and Urban Planning, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; zahriah@usk.ac.id (Z.Z.)
[5] Department of Chemical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; suhendrayatna@usk.ac.id (S.S.)
[6] Department of Medical Chemistry/Biochemistry, Faculty of Medicine, Lambung Mangkurat University, Banjarbaru 70124, Indonesia; ekoantioxidant@gmail.com (E.S.)
[7] Department of Pharmacy, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia; Department of Chemistry, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia; khairankhairan@usk.ac.id (K.K.)
[8] Centre for Technology in Water and Wastewater, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo 2007 NSW Australia; Fitranto.Kusumo@student.uts.edu.au (F.K.)
[9] Department of Chemical Engineering, Universitas Riau, Pekanbaru 28293, Indonesia; zuchra.helwani@lecturer.unri.ac.id (Z.H.)
[10] Faculty of Chemical & Process Engineering Technology, Universiti Malaysia Pahang, Lebuhraya Persiaran Tun Khalil Yaakob, 26300 Gambang, Kuantan, Pahang, Malaysia; sunarti@ump.edu.my (S.A.R.)

***** Correspondence: trizkynoviandy@gmail.com

Article History

Received 16 September 2023
Revised 28 October 2023
Accepted 1 November 2023
Available Online 6 November 2023

Keywords:
Air quality index
Artificial intelligence
Data analysis
Pollutant

Abstract

Urban areas worldwide grapple with environmental challenges, notably air pollution. DKI Jakarta, Indonesia's capital city, is emblematic of this struggle, where rapid urbanization contributes to increased pollutants. This study employed the CatBoost machine learning algorithm, known for its resistance to overfitting and capability to handle missing data, to predict urban air quality based on pollutant levels from 2010 to 2021. The dataset, sourced from Jakarta's air quality monitoring stations, includes pollutants such as PM_{10}, SO_2, CO, O_3, and NO_2. After preprocessing, we used 80% of the data for training and 20% for testing. The model displayed high accuracy (0.9781), precision (0.9722), and recall (0.9728). The feature importance chart revealed O_3 (Ozone) as the top influencer of air quality predictions, followed by PM_{10}. Our findings highlight the dominant pollutants affecting urban air quality in Jakarta, Indonesia and emphasizing the need for targeted strategies to reduce their concentrations and ensure a cleaner and healthier urban environment.

1. Introduction

Urban areas, defined by their busy streets, high-rise buildings, and varied populations, are hubs of human activity and innovation [1]. They play a pivotal role in driving economic development, fostering cultural interactions, and advancing technology. Yet, alongside their vibrant facade, these areas face intricate challenges that jeopardize the environment and the health of their inhabitants [2].

Rapid urbanization in modern cities has significant environmental consequences. As cities grow, energy use, waste production, and strains on transportation all increase. This urban expansion creates a complex web of environmental issues. Notably, air pollution emerges as a predominant and damaging concern, directly impacting the health and living standards of city residents [3].

One of the major urban areas dealing with complex urbanization issues is DKI Jakarta, the capital city of Indonesia [4]. DKI Jakarta is a bustling metropolis with population of over 10 million people [5]. However, rapid urbanization bringing about various issues [6]. Among these, air pollution has emerged as a significant concern. The city's rapid growth has led to increased vehicular traffic and industrial activities, resulting in higher levels of air pollution [7]. This pollution includes the release of harmful airborne particles and pollutants such as particulate matter (PM_{10}), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and nitrogen dioxide (NO_2) poses serious health risks to urban populations, leading to respiratory problems, cardiovascular diseases, and even premature mortality [8, 9].

In recent years, researchers have increasingly utilized machine learning to address urban air pollution challenges. Machine learning is a subset of artificial intelligence that allows systems to learn from data, rather than through explicit programming [10–14]. By recognizing patterns and making decisions based on provided data, it offers the potential for sophisticated data analysis and prediction [15–17]. This approach enhances air quality monitoring systems by leveraging machine learning algorithms to analyze extensive data from sources such as monitoring stations.

Machine learning algorithms provide real-time insights into air quality conditions, enabling the identification of pollution hotspots, prediction of air quality fluctuations, and assessment of pollution control measures' effectiveness [18–20]. However, many previous studies have not fully utilized the capabilities of modern machine learning techniques. Instead of leveraging more advanced algorithms that can handle the complexities of urban pollution data, many have used older, simpler models, potentially missing out on more accurate and insightful predictions.

Air quality prediction accuracy could be enhanced by using advanced machine learning algorithms like CatBoost [21]. CatBoost is a type of gradient-boosting algorithm adept at working with complex, high-dimensional data sets like those used for modeling urban air quality [22]. Key advantages of CatBoost include its robustness against overfitting and its built-in tools for feature importance evaluation, which can be incredibly useful for understanding the relative impact of different pollutants on the predictive model [23–25]. This allows researchers and policymakers to identify the most significant contributors to air quality to help in targeting specific pollutants for mitigation efforts to improve air quality more effectively.

Our study aims to address the gaps left by previous efforts in tackling urban air pollution using machine learning. Recognizing that the complexity of air quality dynamics in metropolises like DKI Jakarta demands sophisticated analytical tools, we propose the application of CatBoost, to provide a more granular and accurate prediction of air quality. Our goal is to not only enhance the accuracy of air quality classification but also to deepen our understanding of pollution patterns and the efficacy of various mitigation strategies.

2. Materials and Methods

2.1. Datasets

The data used in this study is the Air Pollution Standard Index data measured from five air quality monitoring stations located in the DKI Jakarta Province, Indonesia, from 2010 to 2021. This data can be downloaded from the Jakarta Open Data website [26]. The five air quality monitoring stations are situated at Bunderan HI (Central Jakarta), Kelapa Gading (North Jakarta), Jagakarsa (South Jakarta), Lubang Buaya (East Jakarta), and Kebon Jeruk (West Jakarta). The data consists of measurements for five pollutants, namely PM_{10}, SO_2, CO, O_3, and NO_2. Further details regarding these pollutants can be found in Table 1. The dataset also includes classifications into five distinct air quality categories: 'good', 'moderate', 'unhealthy', 'very unhealthy', and 'dangerous', providing a comprehensive view of air quality levels.

2.2. Data Preprocessing

Initially, our dataset consisted of 21,915 samples. Samples with missing values were excluded, resulting in a total of 17,115 samples. Notably, there was a significant class imbalance observed. To address this, the "dangerous" and "very unhealthy" classes were merged

Table 1. Pollutants and their impacts.

Pollutant	Concentration	Explanation	Health Effects
PM_{10} (Particulate Matter 10)	µg/m³	Tiny particles from sources like construction, industry, and natural dust.	Respiratory issues
SO_2 (Sulfur Dioxide)	ppb	A colorless gas from burning fuels, industrial processes, and volcanoes.	Respiratory irritation
CO (Carbon Monoxide)	ppm	Odorless gas from incomplete fuel combustion, especially in vehicles.	Reduced oxygen, headaches
O_3 (Ozone)	ppb	Ground-level ozone, formed by pollutants reacting with sunlight.	Respiratory problems, plant damage
NO_2 (Nitrogen Dioxide)	ppb	Reddish gas primarily from vehicle emissions and combustion processes.	Respiratory problems, lung impairment

Table 2. Descriptive statistics of the dataset.

Pollutant	Mean	SD	Min	Q1	Q2	Q3	Max
PM_{10}	53.32	18.69	2.00	42.00	55.00	65.00	179.00
SO_2	18.18	12.73	0.00	9.00	16.00	25.00	112.00
CO	20.98	12.42	1.00	12.00	18.00	27.00	135.00
O_3	65.67	37.22	3.00	40.00	61.00	83.00	314.00
NO_2	12.50	8.89	1.00	7.00	11.00	16.00	148.00

into a single "unhealthy" class [27]. Consequently, the distribution of the final classes is as follows: "unhealthy" with 2,581 samples, "moderate" with 11,624 samples, and "good" with 2,910 samples.

To prepare the data for model training and evaluation, we split the dataset into training and testing sets. Specifically, 80% of the samples (13,692 samples) were allocated to the training set, and the remaining 20% (3,423 samples) were reserved for the testing set. The rationale behind using an 80-20 split for the training and testing sets is to ensure the model is trained on a sufficiently large and diverse portion of the data so that it can learn the patterns well [28, 29]. The testing set, comprising unseen data not used during training, allows us to evaluate the model's ability to generalize to new data beyond what it has seen before. This helps determine how well the model will perform when deployed in a real-world setting.

2.3. CatBoost

CatBoost is a highly powerful and versatile machine learning algorithm categorized under ensemble learning. It has gained widespread recognition for its remarkable predictive performance and is increasingly popular in various domains owing to its robustness and adaptability. CatBoost excels in handling intricate data relationships. The core technique utilized by CatBoost is gradient boosting, where it iteratively trains weak learners, usually decision trees, to rectify the errors of prior models, culminating in a robust ensemble model [22, 30].

CatBoost boasts several advantages that have contributed to its widespread adoption. To begin with, it incorporates advanced regularization techniques that effectively curb overfitting, enhancing the model's ability to generalize to unseen data. Moreover, CatBoost is proficient at automatically handling missing values during the training process by intelligently determining the best way to distribute samples with incomplete information. This is accomplished through the dynamic optimization of node splits while constructing the decision trees, significantly reducing the necessity for manual preprocessing to impute missing values [24].

2.3. Performance Evaluation

The evaluation encompasses accuracy, precision, recall, and F1-score, all of which provide a comprehensive understanding of the model's performance across different aspects [31, 32]. This array of metrics ensures a thorough evaluation of the proposed approach's strengths and limitations, providing a well-rounded perspective on its overall performance [33–35].

3. Results and Discussion

3.1. Exploratory Data Analysis

To get a comprehensive understanding of the urban air quality dataset, we embarked on an exploratory data analysis. The descriptive statistics of the dataset is presented in Table 2, which provided a snapshot of the distribution and spread of different pollutants. The average PM_{10} concentration was 53.32 µg/m³ with a standard deviation of 18.69 µg/m³, indicating values that range between 2.00 and 179.00 µg/m³. SO_2 had an average concentration of 18.18 ppb, with a standard

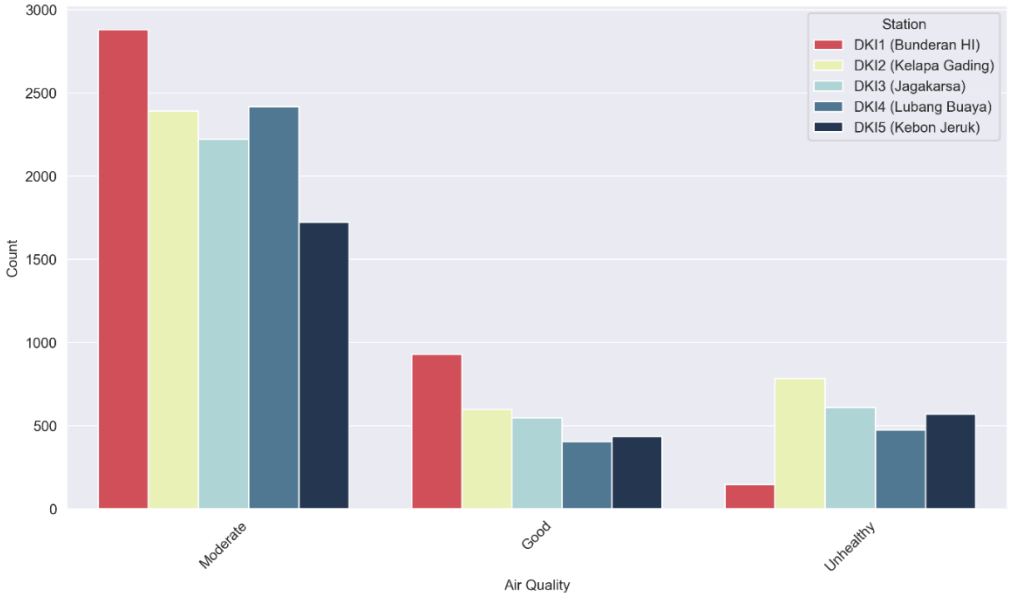

Figure 1. Frequency of different air quality classifications across multiple monitoring stations.

deviation of 12.73 ppb, reflecting a range of values from 0.00 to 112.00 ppb. The average CO concentration was 20.98 ppm, accompanied by a standard deviation of 12.42 ppm, with values spanning from 1.00 to 135.00 ppm. O_3 had a marked average value of 65.67 ppb, with a notably high standard deviation of 37.22 ppb, signifying considerable fluctuations that ranged from 3.00 to 314.00 ppb. This substantial variance in O_3 concentrations, as indicated by the high standard deviation, suggests a wide disparity in readings over time or across different monitoring locations, potentially due to diurnal and seasonal changes, varying weather conditions, traffic volume, and industrial emissions that can rapidly alter O_3 formation and depletion. Lastly, NO_2 presented a mean concentration of 12.50 ppb, a standard deviation of 8.89 ppb, and values that ranged from 1.00 to 148.00 ppb. These statistics provide a foundation for understanding the patterns and variations in air quality data, which are important for environmental and health assessments.

We visualized the frequency of different air quality classifications across multiple monitoring stations (Figure 1). Our analysis of the count data revealed several key findings. The "Moderate" air quality seems to dominate most stations, with the highest count observed in the station DKI1 (Bunderan HI). This is closely followed by stations DKI2 (Kelapa Gading) and DKI3 (Jagakarsa), while DKI4 (Lubang Buaya) and DKI5 (Kebon Jeruk) show slightly lower counts but still remain significant.

For the "Good" air quality classification, DKI3 (Jagakarsa) leads with the highest count, albeit lesser than its "Moderate" count. DKI2 (Kelapa Gading) follows closely, with DKI1 (Bunderan HI) and DKI5 (Kebon Jeruk) displaying a similar pattern of lesser counts. Intriguingly, DKI4 (Lubang Buaya) has a considerably reduced number of "Good" air quality days compared to other stations.

The "Unhealthy" air quality classification, although less frequent compared to the other classifications, still presents considerable differences among the stations. DKI2 (Kelapa Gading) stands out with the most instances, followed by DKI3 (Jagakarsa) and DKI5 (Kebon Jeruk). Both DKI1 (Bunderan HI) and DKI4 (Lubang Buaya) depict relatively fewer occurrences of "Unhealthy" air quality days.

Overall, while most air quality readings across all stations predominantly fall within the "Moderate" range, there are distinct disparities in the distribution of "Good" and "Unhealthy" classifications among different stations. However, it's important to note that the classification counts alone do not necessarily indicate the overall pollution level of a region because our data has been curated to exclude missing values for consistency and reliability in modeling.

3.2. Classification Performance

To achieve our objective of classifying air quality, we utilized the CatBoost algorithm to develop a predictive model. The performance metrics of this model are showcased in Table 3. Impressively, the model exhibits a high accuracy rate of 0.9781, indicating that a significant majority of its predictions align with the true classifications. In terms of precision, a value of 0.9722 suggests that most of the predictions labeled as a specific class are indeed correctly classified. Similarly, the recall score of 0.9728 reflects that the model successfully

Table 3. Model performance on the testing set.

Metrics	Value
Accuracy	97.81%
Precision	97.22%
Recall	97.28%
F1-score	97.24%

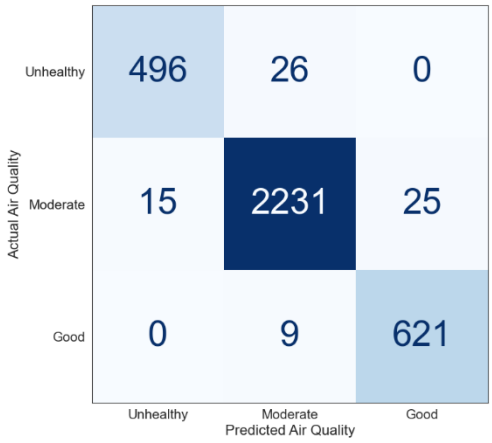

Figure 2. Confusion matrix of the testing set.

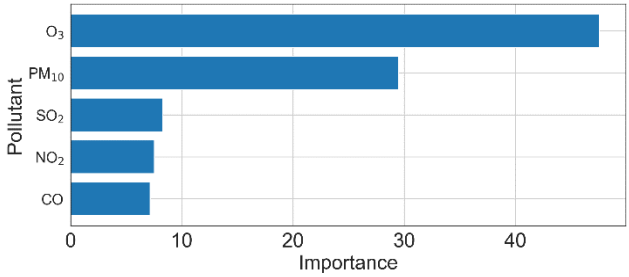

Figure 3. Feature importance of the model.

identifies a large proportion of actual positive instances for each class. Further, the F1 Score, an essential metric that harmonizes the balance between precision and recall, stands at 0.9724, underscoring the model's robust and consistent performance in both aspects.

Further elaborating on the model's performance, the confusion matrix offers a detailed view of its classification results (Figure 2). For air quality labeled as "Unhealthy," the model has made 496 correct predictions, but 26 instances were misclassified as "Moderate" while none were mislabeled as "Good." When considering "Moderate" air quality, the model boasts a robust 2231 accurate predictions. However, there were slight misclassifications with 15 instances predicted as "Unhealthy" and 25 instances as "Good." For instances that were genuinely "Good" in air quality, the model showcased impressive precision by correctly classifying 621 of them. There were only 9 instances that were mistakenly categorized as "Moderate," and the model showed perfect distinction by not misclassifying any "Good" instances as "Unhealthy."

The feature importance chart derived from the CatBoost model provides insightful perspectives into which factors most influence the prediction of air quality (Figure 3). Topping the list, the concentration of O_3 emerges as the most influential feature. Its prominence in the model underscores its critical role in determining air quality, as indicated by its substantial lead in importance compared to other factors.

PM_{10} is another significant contributor. While not as dominant as O_3, its presence is undeniably pivotal in the model's decision-making process. This aligns with real-world observations, as PM_{10} particles can be inhaled easily and are often associated with adverse health effects. SO_2 and NO_2 present a comparable level of importance, suggesting that both play a nearly equivalent role in influencing the air quality predictions. Lastly, CO appears as the least influential among the considered features. While it does have an impact, its role in this specific model's air quality predictions seems to be less pronounced than the other factors.

The results of this study have demonstrated an accurate and precise way to classify urban air quality based on pollutant levels. The model's high accuracy, precision, and recall highlight its potential real-world use for monitoring, predicting, and managing air quality. The finding that O_3 is an important factor influencing air quality predictions shows policymakers must prioritize managing O_3 levels in cities. Given the health risks, the study emphasizes the need to continuously monitor and reduce PM_{10} levels. Additionally, as machine learning provides more actionable insights, urban planners and policymakers have a more powerful tool beyond traditional monitoring. This enhances preemptively controlling air quality degradation and informing policies for healthier cities. In short, combining advanced machine learning with environmental monitoring can greatly boost our efforts to ensure sustainable, livable urban spaces moving forward.

4. Conclusions

Our study demonstrated notable accuracy in classifying urban air quality, particularly in DKI Jakarta, Indonesia, with O_3 and PM_{10} emerging as significant influencing factors. However, the research was geographically limited to Jakarta and only considered five major pollutants, omitting other potential influential factors like meteorological data. There was also a noted class imbalance which might affect predictions. Future research avenues should encompass a broader geographical scope, incorporate additional variables for

a holistic prediction model, and examine the potential of real-time IoT sensor data and other machine learning algorithms to further urban environmental understanding and solutions.

Author Contributions: Conceptualization, G.M.I., T.R.N., A.M. and E.S.; methodology, G.M.I., T.R.N. and Z.Z.; software, T.R.N. and A.M.; validation, S.S., K.K., F.K. and S.A.R.; formal analysis, G.M.I. and Z.Z.; investigation, G.M.I., S.S. and K.K.; resources, F.K., Z.H. and S.A.R.; data curation, E.S. and Z.H.; writing—original draft preparation, G.M.I., T.R.N. and Z.Z.; writing—review and editing, S.S., E.S. and K.K.; visualization, A.M. and S.A.R.; supervision, T.R.N.; project administration, T.R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset utilized for this study is publicly accessible and can be obtained from the Jakarta Open Data website.

Acknowledgments: The authors express their profound gratitude to their respective institutions and universities for their unwavering support and invaluable contributions throughout the course of this study.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Collier, C. G. (2006). The impact of urban areas on weather, *Quarterly Journal of the Royal Meteorological Society*, Vol. 132, No. 614, 1–25. doi:10.1256/qj.05.199.
2. Pateman, T. (2011). Rural and urban areas: comparing lives using rural/urban classifications, *Regional Trends*, Vol. 43, No. 1, 11–86. doi:10.1057/rt.2011.2.
3. Wang, S., Gao, S., Li, S., and Feng, K. (2020). Strategizing the relation between urbanization and air pollution: Empirical evidence from global countries, *Journal of Cleaner Production*, Vol. 243, 118615.
4. Murakami, A., Kurihara, S., Harashina, K., and Zain, A. M. (2017). Features of Urbanization and Changes in the Thermal Environment in Jakarta, Indonesia, *Sustainable Landscape Planning in Selected Urban Regions*, 61–71.
5. Martinez, R., and Masron, I. N. (2020). Jakarta: A city of cities, *Cities*, Vol. 106, 102868.
6. Idroes, G. M., Hardi, I., Nasir, M., Gunawan, E., Maulidar, P., and Maulana, A. R. R. (2023). Natural Disasters and Economic Growth in Indonesia, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 33–39. doi:10.60084/eje.v1i1.55.
7. Lu, J., Li, B., Li, H., and Al-Barakani, A. (2021). Expansion of city scale, traffic modes, traffic congestion, and air pollution, *Cities*, Vol. 108, 102974.
8. Suh, H. H., Bahadori, T., Vallarino, J., and Spengler, J. D. (2000). Criteria air pollutants and toxic air pollutants., *Environmental Health Perspectives*, Vol. 108, No. suppl 4, 625–633.
9. Domingo, J. L., and Rovira, J. (2020). Effects of air pollutants on the transmission and severity of respiratory viral infections, *Environmental Research*, Vol. 187, 109650.
10. Noviandy, T. R., Maulana, A., Idroes, G. M., Emran, T. B., Tallei, T. E., Helwani, Z., and Idroes, R. (2023). Ensemble Machine Learning Approach for Quantitative Structure Activity Relationship Based Drug Discovery: A Review, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 32–41. doi:10.60084/ijds.v1i1.91.
11. Maulana, A., Noviandy, T. R., Sasmita, N. R., Paristiowati, M., Suhendra, R., Yandri, E., Satrio, J., and Idroes, R. (2023). Optimizing University Admissions: A Machine Learning Perspective, *Journal of Educational Management and Learning*, Vol. 1, No. 1, 1–7. doi:10.60084/jeml.v1i1.46.
12. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.
13. Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., and Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–7. doi:10.60084/ijds.v1i1.72.
14. Iffaty, A., Salsabila, A., Rafiqhi, A. A., Suhendra, R., Yusuf, M., and Sasmita, N. R. (2023). Enhancing Water Quality Assessment in Indonesia Through Digital Image Processing and Machine Learning, *Grimsa Journal of Science Engineering and Technology*, Vol. 1, No. 1, 1–7.
15. Mahesh, B. (2020). Machine learning algorithms-a review, *International Journal of Science and Research (IJSR).[Internet]*, Vol. 9, No. 1, 381–386.
16. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.
17. Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T. R., Sasmita, N. R., Subianto, M., Earlia, N., Niode, N. J., and Idroes, R. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification, *Heca Journal of Applied Sciences*, Vol. 1, No. 2, 54–61. doi:10.60084/hjas.v1i2.85.
18. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., and Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California, *Complexity*, Vol. 2020, 1–23. doi:10.1155/2020/8049504.
19. Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., and Harrison, R. M. (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique, *Atmospheric Chemistry and Physics*, Vol. 19, No. 17, 11303–11314. doi:10.5194/acp-19-11303-2019.
20. Masih, A. (2019). Machine learning algorithms in air quality modeling, *Global Journal of Environmental Science and Management*, Vol. 5, No. 4, 515–534. doi:10.22034/GJESM.2019.04.10.
21. Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., and Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis, *Journal of Environmental and Public Health*, Vol. 2023, 1–26. doi:10.1155/2023/4916267.
22. Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support, *ArXiv Preprint ArXiv:1810.11363*.
23. Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., and Arfi, W. Ben. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction, *Technological Forecasting and Social Change*, Vol. 166, 120658. doi:10.1016/j.techfore.2021.120658.
24. Dhananjay, B., and Sivaraman, J. (2021). Analysis and classification of heart rate using CatBoost feature ranking model, *Biomedical Signal Processing and Control*, Vol. 68, 102610. doi:10.1016/j.bspc.2021.102610.

25. Al-Sarem, M., Saeed, F., Boulila, W., Emara, A. H., Al-Mohaimeed, M., and Errais, M. (2021). Feature Selection and Classification Using CatBoost Method for Improving the Performance of Predicting Parkinson's Disease, 189–199. doi:10.1007/978-981-15-6048-4_17.

26. Jakarta Open Data. (2021). Indeks Standar Pencemaran Udara (ISPA), from https://data.jakarta.go.id/dataset/?q=Indeks+Standar+Pencemaran+Udara+&sort=1.

27. Hamami, F., and Dahlan, I. A. (2022). Air Quality Classification in Urban Environment using Machine Learning Approach, *IOP Conference Series: Earth and Environmental Science*, Vol. 986, No. 1, 012004. doi:10.1088/1755-1315/986/1/012004.

28. Joseph, V. R. (2022). Optimal ratio for data splitting, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 15, No. 4, 531–538. doi:10.1002/sam.11583.

29. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., Sasmita, N. R., Muslem, M., Idroes, G. M., Kemala, P., and Irvanizam, I. (2021). Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index, *International Review on Modelling and Simulations (IREMOS)*, Vol. 14, No. 2, 137. doi:10.15866/iremos.v14i2.20460.

30. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, Vol. 31.

31. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.

32. Maulana, A., Noviandy, T. R., Idroes, R., Sasmita, N. R., Suhendra, R., and Irvanizam, I. (2020). Prediction of Kovats Retention Indices for Fragrance and Flavor using Artificial Neural Network, *Proceedings of the International Conference on Electrical Engineering and Informatics* (Vol. 2020-Octob). doi:10.1109/ICELTICs50595.2020.9315391.

33. Carvalho, D. V, Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics, *Electronics*, Vol. 8, No. 8, 832.

34. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.

35. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35. doi:10.60084/ijma.v1i1.78.