Original Article

Check for updates

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

Infolitika Journal of Data Science

Vol. 2, No. 1, 2024

Decision Tree versus k-NN: A Performance Comparison for Air Quality Classification in Indonesia

Novi Reandy Sasmita [1,*], Siti Ramadeska [1], Zurnila Marli Kesuma [1], Teuku Rizky Noviandy [2], Aga Maulana [2], Mhd Khairul [1] and Rivansyah Suhendra [3]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; novireandys@usk.ac.id (N.R.S.); sitiramadeska01@gmail.com (S.R.); zurnila@usk.ac.id (Z.M.K); mhd_khairu@mhs.usk.ac.id (M.K.)
[2] Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23771, Indonesia; trizkynoviandy@gmail.com (T.R.N.)
[3] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; agamaulana@usk.ac.id (A.M.)
[4] Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, West Aceh 23615, Indonesia; rivansyahsuhendra@utu.ac.id (R.S.)

***** Correspondence: novireandys@usk.ac.id

Article History

Received 6 March 2024
Revised 5 May 2024
Accepted 13 May 2024
Available Online 18 May 2024

Keywords:
Air quality
Classification
Decision Tree
K-Nearest Neighbor

Abstract

Air quality can affect human health, the environment, and the sustainability of ecosystems, so efforts are needed to monitor and control air quality. The Plume Air Quality Index (PAQI) is one of the indices to measure and determine the level of air quality. In measuring the accuracy of the air quality level, it is necessary to do the right classification. Some previous studies have conducted classification analysis using the decision tree and K-Nearest Neighbor (k-NN) methods, but only evaluated using accuracy values. Therefore, this study uses both methods to evaluate the results of air quality level classification not only with accuracy but also with precision, recall, and F1-score. Secondary data of pollutant concentration values and PAQI categories based on particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), and ozone (O_3) derived from Plume Labs for 33 provincial capitals in Indonesia in the time period from July 1 to December 31, 2022, were used in this study. From the results of comparing the performance of the two methods, it is found that the decision tree has a greater performance value than the performance value of k-NN. The decision tree performance values for accuracy, precision, recall and F1-score are 90.67%, 90.61%, 90.67%, and 90.63%, respectively. So, it can be concluded that the decision tree performs better than k-NN in classifying PAQI categories with better overall evaluation metric values.

1. Introduction

Air quality is decreasing along with population growth and the development of technology and science, which can lead to air pollution. Air quality maintenance must be done to maintain human health and protect other living things [1]. However, in reality, air quality is not always clean. Air pollution can originate from a diverse range of origins, including natural occurrences like wind-driven dust and volcanic eruptions, as well as human activities such as emissions from power plants, factories, vehicles, and the incineration of forests or any form of waste [2, 3].

Based on the World Health Organization (2021), air pollution does not rule out the possibility of many people developing stroke, heart disease, lung cancer, and chronic and acute respiratory diseases, such as asthma. Indonesia is ranked first in the region of Southeast Asia

and 17th in the world as the most polluted country based on the calculation of the annual average $PM_{2.5}$ particulate concentration (µg/m3) in 2021. The problem of air quality degradation is caused by the concentration of each pollutant. Each pollutant does not have the same impact on human health at the same concentration. Currently, countries in Europe, the United States, and Asia have published air quality data, but do not use the same scale for Air Quality Index (AQI) categorization. It is not easy to compare air quality levels on a global scale because each AQI calculation system is based on different pollutants, thresholds, methods, and a number of categories. Therefore, Plume Labs, an environmental technology company that provides air monitoring services, created a new air quality measurement system, the Plume Air Quality Index (PAQI) [4]. The more accurate the information provided, the more it will help the government design policies and programs to maintain air quality and human health. One way that can be done to find out how good the classification produced by Plume Labs is in providing air quality information is to process data with a data mining and machine learning approach [5].

One of the techniques found in data mining is classification. Classification is often used in predicting classes on certain attributes [6]. Classification is categorized into five groups rooted in mathematical principles: statistical-based, distance-based, decision tree-based, neural network-based, and rule-based approaches [7]. There are many types of algorithms in classification techniques, and this study uses the decision tree and K-Nearest Neighbor (k-NN) algorithms.

In recent years, multiple studies have examined the classification performance of Decision Tree and k-NN methods. Daldiri and Fitriati [8] compared k-NN and decision tree for classifying breast cancer and reported the highest performance achieved by a decision tree with an accuracy of 93.3%. Krishna and Rama Parvathy [9] compared k-NN and decision tree accuracy prediction of medical insurance and found that k-NN is the more promising result with 87.4% accuracy.

There are several reasons for comparing decision tree and k-NN methods. The main reason for comparing them is because they are both commonly used algorithms in data modeling and classification. Decision trees and k-NN are often contrasted in machine learning. A decision tree makes decisions based on rules and separates data into different groups based on relevant features, offering an advantage in interpretability due to its easy-to-understand tree structure [10]. On the other hand, k-NN makes decisions based on the distance between the data to be classified and the existing training data, which lacks a clear structure and makes interpretation difficult [11]. While numerous individual studies have explored the application of decision tree and k-NN classification techniques, there is currently no ongoing study employing PAQI data in Indonesia, accompanied by model assessment utilizing performance metrics such as accuracy, recall, precision, and F1-score.

Each algorithm possesses its unique set of strengths and weaknesses when it comes to data classification, so the best classification performance results will vary in each study [12]. Both performances are compared with the aim of finding out which algorithm is the best to apply in this study.

2. Materials and Methods

2.1. *Data Description*

This study uses secondary data from Plume Labs (https://air.plumelabs.com/en/) regarding pollutant concentrations with concentration measurement units (µg/m3) and PAQI categories from 33 provincial capitals in Indonesia available on the website for the period 1 July to 31 December 2022 or as many as 184 days. PAQI category is used as a label and pollutant concentration is used as a feature.

PAQI thresholds are based on WHO recommendations that take into account four types of pollutants harmful to the environment and human health: particulate matter PM_{10}, $PM_{2.5}$, nitrogen dioxide NO_2, and ozone O_3. The PAQI consists of five pollution levels with thresholds for each category, namely "Fresh Air" (0-50), "Moderate Pollution" (51-100), "High Pollution" (101-150), "Very High Pollution" (151-200), and "Excessive Pollution" (>200). This range of values reflects air quality levels from very good to very hazardous, providing important insights for decision-making in efforts to improve air quality and public health [4].

2.2. Classification Algorithms, Data Sharing and Performance Evaluation

In this study, the algorithms used are decision tree and k-NN. There are several data division partitions used in this study with a ratio of 70:30, 80:20, and 90:10, each implemented to find the ratio with the best performance.

Model evaluation in this study uses the confusion matrix method which can represent information on the comparison of prediction results with actual conditions [13, 14]. Performance is measured based on a combination of accuracy, precision, recall, and F1-score [15]. The equation for each of these calculations are shown in Equation 1-4.

Table 1. Concentration and dispersion of pollutant data (n=184 days).

No.	Features	Descriptive Statistics						
		Mean	Std	Min	Q1	Q2	Q3	Max
1	$PM_{2.5}$	11.19	18.42	0	0.00	4.00	14.00	146
2	PM_{10}	17.61	26.75	0	1.00	8.00	22.00	211
3	NO_2	12.93	26.57	0	1.00	5.00	12.00	278
4	O_3	34.15	14.25	0	24.75	32.00	42.00	124

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} x\ 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} x\ 100\% \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)} x\ 100\% \quad (3)$$

$$F1 - score = 2\ x\ \frac{(Precision\ x\ Recall)}{(Precision\ +\ Recall)} x\ 100\% \quad (4)$$

2.3. Data Analysis

In this study, statistics is the primary method used to generalize the results [16–19]. Data analysis used Python programming language software version 3.10.0 with Jupyter Notebook editor version 6.5.2. Descriptive analysis was used to calculate data concentration and data distribution to summarize and describe the characteristics of a dataset [14, 20, 21]. The statistical method used in this study consists of frequency as part of descriptive statistical analysis [18, 20, 22]. Then, the visualization of PAQI distribution in each provincial capital in Indonesia is displayed in the form of a thematic map.

In the initial stage of analysis for machine learning, the data is divided into two parts, namely training data and testing data, with three division ratios, namely 70:30, 80:20, and 90:10. This division aims to obtain optimal classification results based on different ratios.

The classification process with decision tree and k-NN on training data begins with hyperparameter tuning. In the decision tree, hyperparameters are used such as criterion value, max. depth, min samples leaf, and min split. Meanwhile, the number of neighbors (k), distance metric, and distance weight are used as hyperparameters in k-NN. Both hyperparameter tuning use the grid search method with 10-fold cross-validation to find the optimal combination that produces the highest accuracy value and form a model based on the combination.

After the model is formed, performance evaluation is conducted for each data partition ratio. The best-performing partition is selected based on the results of training data performance, and the best algorithm is selected based on the comparison of testing data performance. The results are used as the conclusion of the research.

2.4. Data Preprocessing

Data standardization is performed using z-score normalization to overcome the difference in the range of values for each feature. Furthermore, we apply the random oversampling technique to balance the dataset. This technique is important when dealing with imbalanced data, where one class significantly outnumbers the other(s) [19, 23, 24].

3. Results and Discussion

The results of data centering and dispersion analysis for pollutant concentrations in Indonesia are shown in Table 1. Based on the table, the highest concentration is found in NO_2 pollutants with a value of 278 μg/m³, while the lowest concentration value is found in all pollutants, which is 0 μg/m³.

For air conditions in Indonesia, of the seven PAQI categories used by Plume Labs, there are only five categories, namely fresh air, moderate pollution, high pollution, very high pollution, and excessive pollution. The "Fresh air" category dominates with a total of 3,673 observations (60.49%). Meanwhile, the "Moderate" category accounts for 1,549 observations (25.51%). The "High" category has 551 observations (9.07%), followed by "Very high" with 244 observations (4.02%). Finally, the "Excessive" category is the least category with only 55 observations (0.91%). This data distribution provides an overview of the distribution of PAQI categories in the data sample used in the analysis. Pollutant distribution based on provincial capitals in Indonesia can be seen in the boxplot shown in Figure 1.

Compared to other provincial capitals, Jakarta has a very high concentration of pollutants, particularly for $PM_{2.5}$, PM_{10}, and NO_2. In contrast, all of the boxes from the provinces are practically in the same place in the O_3 boxplot. Additionally, it is clear that each pollutant has outliers, but Jakarta's NO_2 levels are particularly severe.

Furthermore, the distribution of air quality based on the average PAQI value in each of Indonesia's provincial

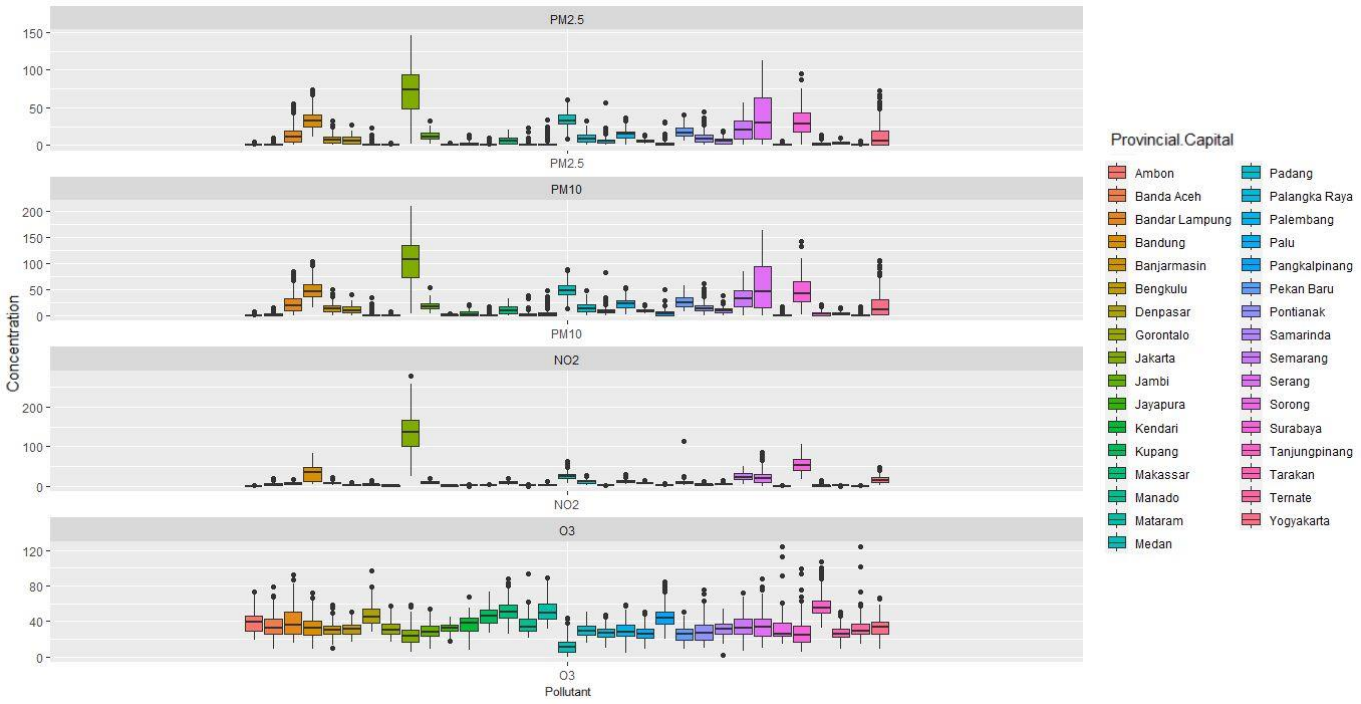

Figure 1. Pollutant distribution based on provincial capitals in Indonesia.

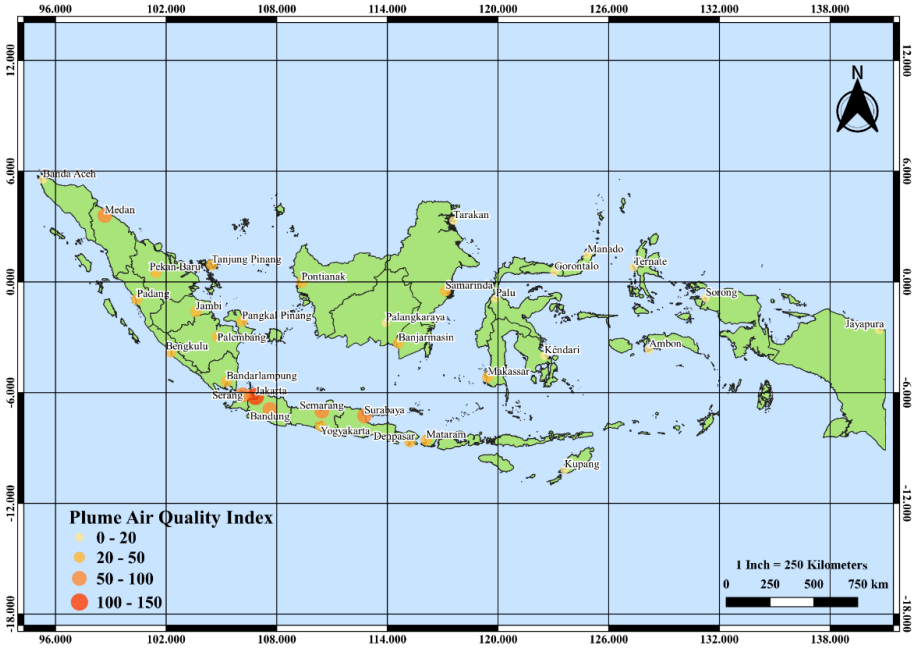

Figure 2. Distribution of PAQI in Indonesia's provincial capitals.

capitals is presented visually in the thematic map in Figure 2. On the map, dots represent provincial capitals, and a small dot with a PAQI value range of 0 - 20 indicates a low average PAQI value, which means a fresh air category. The larger the dot, the higher the PAQI value in that provincial capital. Then, Figure 1 shows that Indonesia is dominated by provincial capitals with average PAQI values in the Moderate pollution category. However, Jakarta City has the highest PAQI value in the very high pollution category, indicated by the largest dot on the map.

A comparison of the amount of data on the target variable before and after data balancing is shown in Table 2. Based on the table, it is known that the number of observations of major and minor classes in the target variable is balanced.

The next stage is the formation of the decision tree model on the training data. The parameters used to control the

Table 2. The number of training data before and after balancing.

Class data	Training data			Random Oversampling Data		
	70%	80%	90%	70%	80%	90%
Fresh air	2576	2941	3308	2576	2941	3308
Moderate	1073	1230	1387	2576	2941	3308
High	388	449	502	2576	2941	3308
Very high	177	195	217	2576	2941	3308
Excessive	36	42	50	2576	2941	3308

Table 3. Performance comparison.

Algorithm	Accuracy	Precision	Recall	F1-score
Decision tree	90.67%	90.61%	90.67%	90.63%
k-NN	84.91%	85.98%	84.91%	85.25%

tree formation process in this study are criterion, min samples split, min samples leaf, and max depth. The tuning results show the use of the Gini criterion with a min samples split value of 2 samples, min samples leaf of 1 sample, and max depth of 5 for all three ratios. With these settings, the model achieved a high level of accuracy, which was about 90.61% for the 70:30 ratio, 89.87% for the 80:20 ratio, and 90.10% for the 90:10 ratio.

Then a decision tree model with optimal parameters on each data partition is built. Then the accuracy, precision, recall, and F1-score values were calculated. The analysis results showed that in the 90:10 partition, the model achieved an accuracy of 90.10%, precision of 90.29%, recall of 90.10%, and F1-score of 90.16%. At 80:20 partition, the accuracy reached 89.87%, precision 90.01%, recall 89.87%, and F1-score 89.92%. While in the 70:30 partition, accuracy reached 90.61%, precision 90.66%, recall 90.61%, and F1-score 90.63%. So that the 70:30 partition provides the best classification results with the highest accuracy, precision, recall, and F1-score compared to other partitions.

The next stage is to evaluate the testing data with the aim of knowing how well the decision tree model has been built to predict accurate prediction results on new data. The analysis results show that this model achieved an accuracy of 90.67%, precision of 90.61%, recall of 90.67%, and F1-score of 90.63%. These performance results show that the decision tree model performs well in classifying plume air quality index data. Then, perform the k-NN algorithm classification. Before the formation of the model, hyperparamter tuning is necessary. In k-NN, there are three hyperparameters: the number of neighbors (k), distance metric, and distance weight. The tuning results show the use of the same distance metric and distance weight for the three data partitions, namely manhattan distance and distance weight, respectively. Then the tuning results for the number of neighbors in the 70:30 partition are 7 neighbors, and the other two partitions are 13 neighbors. With these settings, the model achieves a high level of accuracy, which is about 97.37% for the 70:30 ratio, 97.36% for the 80:20 ratio, and 97.41% for the 90:10 ratio.

The next step is to build a k-NN model with the optimal hyperparameters that have been obtained.The analysis results show that in the 90:10 partition, the model achieved 97.39% accuracy, 97.39% precision, 97.31% recall, and 97.30% F1-score.In the 80:20 partition, the accuracy reached 97.36%, precision 97.44%, recall 97.36%, and F1-score 97.36%.While in the 70:30 partition, the accuracy reached 97.38%, precision 97.44%, recall 97.38%, and F1-score 97.37%.The highest value for each performance calculation metric is found in the 70% training and 30% testing data partition.After obtaining evaluation results on training data, an evaluation will be carried out to measure the performance or performance of the k-NN algorithm on testing data.

The highest performance calculation results generated by data partitioning 70% training data and 30% testing data with the calculation results, namely the accuracy value of 84.91%, precision of 85.98%, recall of 84.91%, and f1-score of 85.25%. This performance value shows that the results are good in classifying PAQI for provincial capitals in Indonesia using k-NN.

Based on the results of the performance values obtained for the decision tree algorithm and k-NN can be seen in Table 3. The results of the performance calculation of the decision tree have a better performance value than k-NN. The results of the decision tree performance calculation, namely the accuracy value of 90.67%, the precision value of 90.61%, the recall value of 90.67%, and the F1-score value of 90.63%. From these results, it can be said that the decision tree algorithm is the best algorithm with a higher performance value compared to k-NN.

This study shows that NO_2 is the pollutant with the highest concentration in Indonesia. This is because NO_2 is a major atmospheric pollutant generated from various anthropogenic sources, especially motor vehicle

emissions, and is associated with population density [25]. High levels of NO_2 are found in and around urban environments [26].

Furthermore, based on the comparison of simulation results, the decision tree is the best algorithm compared to the k-NN algorithm or other machine learning classification algorithms. This is in accordance with previous research conducted by Bilen and Bozkurt [27] who discussed the comparative analysis of several machine and deep learning algorithms to predict the AQI based on data collected from various stations in Kocaeli Province. Researchers used decision tree, k-NN, Naive Bayes, Logistic Regression, SVM, RF, RNN, and LSTM algorithms and found that the most accurate classification was performed by the decision tree algorithm with a maximum accuracy of 94%. Furthermore, previous research mentioned that in classifying air quality in Indonesia the decision tree is a very good method [28].

Furthermore, there are several reasons why decision trees can be better than k-NN [29]. First, the decision tree takes less time in the classification process than k-NN, especially with large datasets, because KNN scans the entire dataset to predict and does not generalize the data first. Secondly, the decision tree is a machine learning algorithm that supports automatic feature interaction. That is, the decision tree can find the relationship between features automatically and generate more complex rules than k-NN. Meanwhile, k-NN only calculates the distance between data points and other data points to determine the closest class. Third, k-NN needs to keep all the training data in memory to compare it with new data points during prediction. While the decision tree does not need to do this activity.

As the size of the training dataset increases, more memory is required to store all the data, which results in slower computation of k-NN classification. This is due to the expansion of the search to the nearest neighbor in the feature space. Furthermore, k-NN is not sensitive to irrelevant data or noise in the dataset. Irrelevant data may affect the classification results in an undesirable way. In addition, k-NN requires the selection of parameter k, which is the number of nearest neighbors to be used in classification. Improper parameter selection can affect the algorithm's performance. A value of k that is too small easily results in overfitting. This is caused by the model focusing too much on specific training data and not generalizing well to new data [30].

Furthermore, the limitation of this research is the time span of the data taken, so it is possible that it cannot cover the optimal conditions of pollutants in Indonesia. This may affect the classification performance of both algorithms.

4. Conclusions

The capital city of Jakarta has the worst air quality in Indonesia based on the average PAQI value in the period July 1 to December 31, 2022, when compared to other provincial capitals. The best classification performance results using both Decision tree and k-NN are generated by 70% data partition for training and 30% for testing. Based on the tuning results of the decision tree hyperparameters, the best parameters are obtained using criterion gini, max depth of 5, min samples leaf of 1 sample, and min samples split of 2 samples. Furthermore, for the k-NN algorithm, the value of k = 11, distance metric manhattan, and distance weight distance are the best hyperparameter tuning results. Furthermore, the results of the Decision Tree performance calculation for accuracy, precision, recall, and F1-score are 90.67%, 90.61%, 90.67%, and 90.63%, respectively. Meanwhile, in the performance of k-NN, the accuracy, precision, recall, and F1-score were 84.91%, 85.98%, 84.91%, and 85.25%, respectively. Therefore, the Decision Tree shows the best performance results in classifying PAQI data in Indonesia because the performance value of decision tree is higher than k-NN for all metrics.

Author Contributions: Conceptualization, N.R.S; methodology, N.R.S, and S.R.; software, S.R; validation, Z.M.K, T.R.N., and A.M.; formal analysis, N.R.S, and S.R.; investigation, R.S.; resources, S.R.; data curation, N.R.S., and A.M.; writing—original draft preparation, N.R.S., and S.R.; writing—review and editing, N.R.S., S.R., and M.K.; visualization, S.R.; supervision, Z.M.K.; project administration, S.R. 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: Data is available upon request to the authors.

Acknowledgments: Thanks to the Plume Labs website for providing free access to data and making valuable contributions in support of this research.

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

References

1. Agista, P. I., Gusdini, N., and Maharani, M. D. D. (2020). Air Quality Analysis with Air Pollution Standard Index (ISPU) and the Distribution of Pollutant Levels in DKI Jakarta Province, *Jurnal SEOI – Fakultas Teknik Universitas Sahid Jakarta*, Vol. 2, No. 2, 39–57.

2. Ahmad, N., Ul-Saufie, A. Z., Shaziayani, W. N., Abidin, A. W. Z., Zulazmi, N. E. S., and Harb, S. M. (2022). Evaluating the Performance of Random Forest and Multiple Linear Regression for Higher Observed PM10 Concentrations, *Israa University Journal of Applied Science*, Vol. 6, No. 1, 72–90. doi:10.52865/WHPM9019.

3. Idroes, G. M., Noviandy, T. R., Maulana, A., Zahriah, Z., Suhendrayatna, S., Suhartono, E., Khairan, K., Kusumo, F., Helwani, Z., and Abd Rahman, S. (2023). Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring, *Leuser Journal of Environmental Studies*, Vol. 1, No. 2, 62–68. doi:10.60084/ljes.v1i2.99.

4. Plume. (2019). *Plume AQI: An Air Quality Index Aligned with Health Recommendations*.

5. Sanmorino, A., Alie, J., Ariati, N., and Wulanda, S. V. (2022). K-NN Based Air Classification as Indicator of the Index of Air Quality in Palembang, *Jurnal Dan Penelitian Teknik Informatika*, Vol. 7, No. 3, 853–859. doi:10.33395/sinkron.v7i3.11469.

6. Idroes, R., Maulana, A., Noviandy, T. R., Suhendra, R., Sasmita, N. R., Lala, A., and Irvanizam. (2020). A Genetic Algorithm to Determine Research Consultation Schedules in Campus Environment, *IOP Conference Series: Materials Science and Engineering*, Vol. 796, 012033. doi:10.1088/1757-899X/796/1/012033.

7. 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, *2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, IEEE, 1–5. doi:10.1109/ICELTICs50595.2020.9315391.

8. Daldiri, Z. F., and Fitriati, D. (2023). Comparison of Breast Cancer Classification Using the Decision Tree ID3 Algorithm and K-Nearest Neighbors Algorithm, *Jurnal Riset Informatika*, Vol. 5, No. 2, 177–186. doi:10.34288/jri.v5i2.406.

9. Krishna, A., and Rama Parvathy, L. (2022). Comparison of Accuracy Prediction of Medical Insurance Using Decision Tree with K-Nearest Neighbour, *Advances in Parallel Computing* (Vol. 0), 493–499. doi:10.3233/APC220070.

10. Pratyusha, M., and Kanimozhi, K. V. (2022). Heart Disease Prediction Using Decision Tree in Comparison with K-Nearest Neighbor to Improve Accuracy, *Advances in Parallel Computing*, Vol. 0, No. 41, 231–236. doi:10.3233/APC220031.

11. Rajaguru, H., and Sannasi Chakravarthy, S. R. (2019). Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer, *Asian Pacific Journal of Cancer Prevention*, Vol. 20, No. 12, 3777–3781. doi:10.31557/APJCP.2019.20.12.3777.

12. Agustia, M., Noviandy, T. R., Maulana, A., Suhendra, R., Muslem, M., Sasmita, N. R., Idroes, G. M., Rahimah, S., Afidh, R. P. F., Subianto, M., Irvanizam, I., and Idroes, R. (2022). Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances, *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, IEEE, 13–18. doi:10.1109/ICELTICs56128.2022.9932124.

13. Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts, Models, Methods, and Algorithms* (3rd ed.), Elsevier ; Morgan Kaufmann.

14. Nadia, Y., Ramli, M., Muslem, Japnur, A. F., Rusyana, A., Idroes, G. M., Suhendra, R., Muhammad, Sasmita, N. R., Tallei, T. E., and Idroes, R. (2019). Simple Combination Method of FTIR Spectroscopy and Chemometrics for Qualitative Identification of Cattle Bones, *IOP Conference Series: Earth and Environmental Science*, Vol. 364, 012040. doi:10.1088/1755-1315/364/1/012040.

15. Chang, V., Bailey, J., Xu, Q. A., and Sun, Z. (2022). Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms, *Neural Computing and Applications*. doi:10.1007/s00521-022-07049-z.

16. Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., and Idroes, R. (2022). The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression, *Journal of Engineering Science and Technology*, Vol. 17, No. 1, 306–326.

17. Sasaki, D., Sofyan, H., Sasmita, N. R., Affan, M., and Nizamuddin, N. (2021). Assessing the Intermediate Function of Local Academic Institutions During the Rehabilitation and Reconstruction of Aceh, Indonesia, *Journal of Disaster Research*, Vol. 16, No. 8, 1265–1273. doi:10.20965/jdr.2021.p1265.

18. Earlia, N., Bulqiah, M., Muslem, M., Karma, T., Suhendra, R., Maulana, A., Amin, M., Sasmita, N. R., Idroes, G. M., and Prakoeswa, C. (2021). Protective Effects of Acehnese Traditionally Fermented Coconut Oil (Pliek U Oil) and its Residue (Pliek U) in Ointment against UV Light Exposure: Studies on Male Wistar Rat Skin (Rattus novergicus), *Sains Malaysiana*, Vol. 50, No. 5, 1285–1295.

19. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., Sasmita, N. R., Muslem, M., Idroes, G. M., and Irvanizam, I. (2019). Retention Index Prediction of Flavor and Fragrance by Multiple Linear Regression and the Genetic Algorithm, *International Review on Modelling and Simulations (IREMOS)*, Vol. 12, No. 6, 373. doi:10.15866/iremos.v12i6.18353.

20. Idroes, R., Husna, I., Muslem, Mahmudi, Rusyana, A., Helwani, Z., Idroes, G. M., Suhendra, R., Yandri, E., Rahimah, S., and Sasmita, N. R. (2019). Analysis of Temperature and Column Variation in Gas Chromatography to Dead Time of Inert Gas and N-Alkane Homologous Series Using Randomized Block Design, *IOP Conference Series: Earth and Environmental Science* (Vol. 364), IOP Publishing, 12020. doi:10.1088/1755-1315/364/1/012020.

21. Azharuddin, A., Sasmita, N. R., Idroes, G. M., Andid, R., Raihan, R., Fadlilah, T., Earlia, N., Ridwan, T., Maya, I., and Farnida, F. (2023). Patient Satisfaction and its Socio-Demographic Correlates in Zainoel Abidin Hospital, Indonesia: A Cross-Sectional Study, *Unnes Journal of Public Health*, Vol. 12, No. 2, 57–67. doi:10.15294/ujph.v12i2.69233.

22. Sofyan, H., Diba, F., Susanti, S. S., Marthoenis, M., Ichsan, I., Sasmita, N. R., Seuring, T., and Vollmer, S. (2023). The State of Diabetes Care and Obstacles to Better Care in Aceh, Indonesia: A Mixed-Methods Study, *BMC Health Services Research*, Vol. 23, No. 1, 271. doi:10.1186/s12913-023-09288-9.

23. He, H., Zhang, W., and Zhang, S. (2018). A Novel Ensemble Method for Credit Scoring: Adaption of Different Imbalance Ratios, *Expert Systems with Applications*, Vol. 98, 105–117. doi:10.1016/j.eswa.2018.01.012.

24. 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.

25. Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., and Di, B. (2018). Satellite-Based Estimates of Daily NO 2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model, *Environmental Science & Technology*, Vol. 52, No. 7, 4180–4189. doi:10.1021/acs.est.7b05669.

26. Beckwith, M., Bates, E., Gillah, A., and Carslaw, N. (2019). NO2 Hotspots: Are We Measuring in the Right Places?, *Atmospheric Environment: X*, Vol. 2, 100025. doi:10.1016/j.aeaoa.2019.100025.

27. Bilen, Z., and Bozkurt, F. (2021). Comparison of Different Machine and Deep Learning Techniques to Predict Air Quality Index: A Case of Kocaeli Province, *2021 29th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 1–4. doi:10.1109/SIU53274.2021.9477936.

28. Eliyati, N., Rahmayani, M., Wijaya, S., Zayanti, D. A., Kresnawati, E. S., and Resti, Y. (2022). Prediction of Air Quality Index Using Decision Tree with Discretization, *Indonesian Journal of Engineering and Science*, Vol. 3, No. 3, 061–067. doi:10.51630/ijes.v3i3.82.

29. Mohanapriya, M., and Lekha, J. (2018). Comparative Study between Decision Tree and KNN of Data Mining Classification Technique, *Journal of Physics: Conference Series*, Vol. 1142, No. 1. doi:10.1088/1742-6596/1142/1/012011.

30. Gou, J., Qiu, W., Yi, Z., Shen, X., Zhan, Y., and Ou, W. (2019). Locality Constrained Representation-Based K-Nearest Neighbor Classification, *Knowledge-Based Systems*, Vol. 167, 38–52. doi:10.1016/j.knosys.2019.01.016.