Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i3.201

Research Article

Predictive Modeling of Air Quality Levels Using Decision Tree Classification: Insights from Environmental and Demographic Factors

I Gede Iwan Sudipa [1,*], Muhammad Habibi [2], Ery Setiyawa Jullev Atmadji [3], Ika Arfiani [4]

[1] Institut Bisnis dan Teknologi Indonesia, Indonesia, iwansudipa@instiki.ac.id
[2] Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia, muhammadhabibi17@gmail.com
[3] Politeknik Negerei Jember, Jember, Indonesia, ery@polije.ac.id
[4] Universitas Ahmad Dahlan, Indonesia, ika.arfiani@tif.uad.ac.id

Correspondence should be addressed to I Gede Iwan Sudipa; iwansudipa@instiki.ac.id

Received 10 October 2024; Accepted 20 December 2024; Published 31 December 2024

Abstract:

Air pollution poses a significant global challenge, adversely impacting public health and environmental sustainability. Understanding the factors influencing air quality is essential for developing effective mitigation strategies. This study aims to analyse key environmental and demographic factors, such as PM2.5 concentration, population density, and proximity to industrial areas, to predict air quality levels using a Decision Tree model. The dataset, comprising 5000 samples, was pre-processed by encoding the target variable and applying Z-score normalization to numerical features. The model was trained on 80% of the data and evaluated on the remaining 20%, achieving an accuracy of 93%. Evaluation metrics, including a classification report and confusion matrix, demonstrated the model's effectiveness in distinguishing between four air quality categories: Good, Moderate, Poor, and Hazardous. PM2.5 emerged as the most critical predictor, followed by demographic and industrial factors. These findings underscore the potential of machine learning models in providing actionable insights for air quality management. The results contribute to public policy by highlighting the need for targeted interventions in high-risk areas and the importance of incorporating environmental data into urban planning. Future work should focus on expanding the feature set and exploring ensemble techniques to further enhance predictive accuracy and robustness.

Keywords: Air Quality, Decision Tree, Environmental Factors, Machine Learning, Public Policy.
Dataset link: https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment

1. Introduction

Air quality has become a critical concern worldwide due to its significant impact on public health and environmental sustainability. Poor air quality not only exacerbates respiratory and cardiovascular diseases but also contributes to climate change and environmental degradation. As urbanization and industrialization accelerate, the need for robust and reliable methods to monitor, predict, and mitigate air pollution has never been more urgent. Governments and organizations globally are formulating stringent policies and regulations to address air quality issues; however, these efforts require accurate data-driven insights to be effective.

Understanding the driving factors behind air pollution is essential for crafting targeted interventions. Various pollutants, such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO), interact with environmental and demographic conditions to influence air quality. The complexity of these interactions poses a significant challenge to traditional analytical methods. Thus, leveraging advanced computational techniques, such as machine learning [1], offers an opportunity to unravel these complexities and make precise predictions about air quality levels.

This research aims to develop a predictive model using Decision Tree classification to categorize air quality levels based on environmental and demographic factors [2]–[4]. By focusing on variables such as temperature, humidity, proximity to industrial zones, and population density, the study seeks to provide a comprehensive understanding of how these factors contribute to varying pollution levels. The scope of the research encompasses not only the development of the model but also its evaluation using established metrics to ensure reliability and applicability in real-world scenarios.

The primary contribution of this study lies in its ability to transform raw environmental data into actionable insights. The Decision Tree model [3], [5]–[7] offers a transparent and interpretable approach to understanding the relative importance of various factors influencing air quality. This understanding is critical for policymakers and urban planners in designing effective strategies to mitigate pollution and improve the quality of life in affected regions.

Furthermore, the insights derived from this research can guide the implementation of early warning systems for hazardous air quality conditions. By identifying high-risk areas and populations, stakeholders can allocate resources more efficiently and prioritize interventions where they are needed the most. The study also contributes to the broader body of literature on machine learning applications in environmental science, demonstrating the potential of data-driven approaches to address pressing societal challenges.

2. Method:

The study adopts a supervised classification approach, leveraging a Decision Tree algorithm to classify air quality into predefined categories: Good, Moderate, Poor, and Hazardous. Decision Trees are chosen due to their interpretability and ability to handle both categorical and numerical data effectively [8], [9]. **Figure 1** show the research workflow involves data pre-processing, model training, and evaluation to ensure accurate and reliable classification of air quality levels [10]–[12].

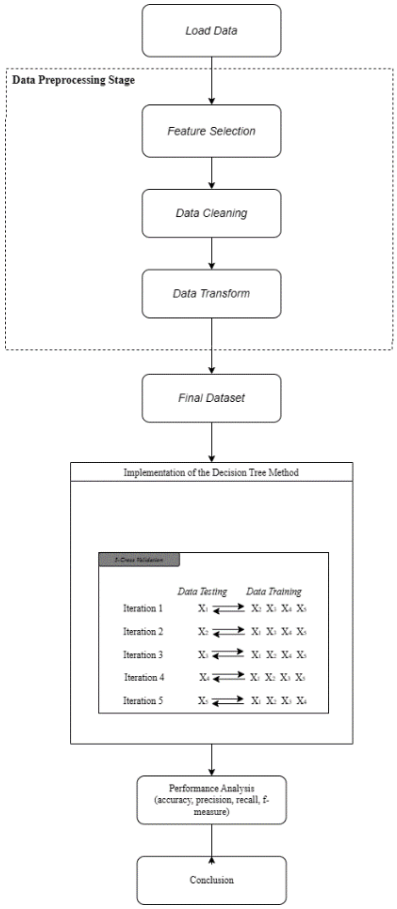

Figure 1. General Research Design Stages

Data Collection Process

The dataset used in this research consists of 5000 samples, capturing critical environmental and demographic factors. The dataset includes the following **Table 1**:

Table 1. Feature Descriptions

Feature	Description	Type
Temperature (°C)	Average temperature of the region	Numerical
Humidity (%)	Relative humidity recorded in the region	Numerical
PM2.5 Concentration (µg/m³)	Fine particulate matter levels	Numerical
PM10 Concentration (µg/m³)	Coarse particulate matter levels	Numerical
NO2 Concentration (ppb)	Nitrogen dioxide levels	Numerical
SO2 Concentration (ppb)	Sulfur dioxide levels	Numerical
CO Concentration (ppm)	Carbon monoxide levels	Numerical
Proximity to Industrial Areas (km)	Distance to the nearest industrial zone	Numerical
Population Density (people/km²)	Number of people per square kilometre in the region	Numerical
Air Quality Levels	Categorical target variable	Categorical

Sampling of the dataset is stratified to ensure proportional representation of each air quality category. This ensures the model is trained on a balanced dataset, mitigating biases and enhancing generalizability.

Data Pre-processing

Pre-processing involves two key steps: encoding and normalization:

1. Encoding: The categorical target variable, `Air Quality Levels`, is encoded into numerical categories as follows:

$$Encoded\ value = \begin{cases} 0 & if\ Good \\ 1 & if\ Moderate \\ 2 & if\ Poor \\ 3 & if\ Hazardous \end{cases} \quad (1)$$

2. Normalization: Normalization is applied to all numerical features to ensure consistent scaling using Z-score normalization [13], [14]. The formula for Z-score normalization is:

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

Where:

z is normalized value

x is original value

μ is mean of the feature

σ is standard deviation of the feature

Normalization ensures that features with larger ranges do not dominate the learning process.

Model Tools

The Decision Tree classifier is implemented using the Python scikit-learn library. Key functionalities of scikit-learn used in the study include:

1. `train_test_split`: To split the dataset into training (80%) and testing (20%) subsets, show in **Figure 2**.

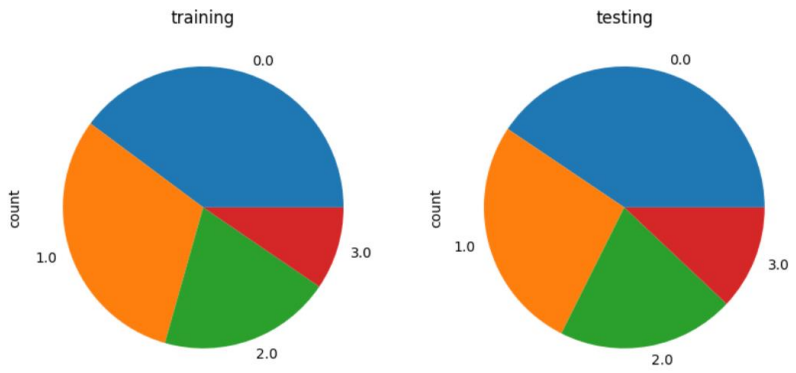

Figure 2. Split Data

2. `DecisionTreeClassifier`: To train and classify air quality levels.
3. `StandardScaler`: To apply Z-score normalization, visualization of normalization show in **Figure 3**.

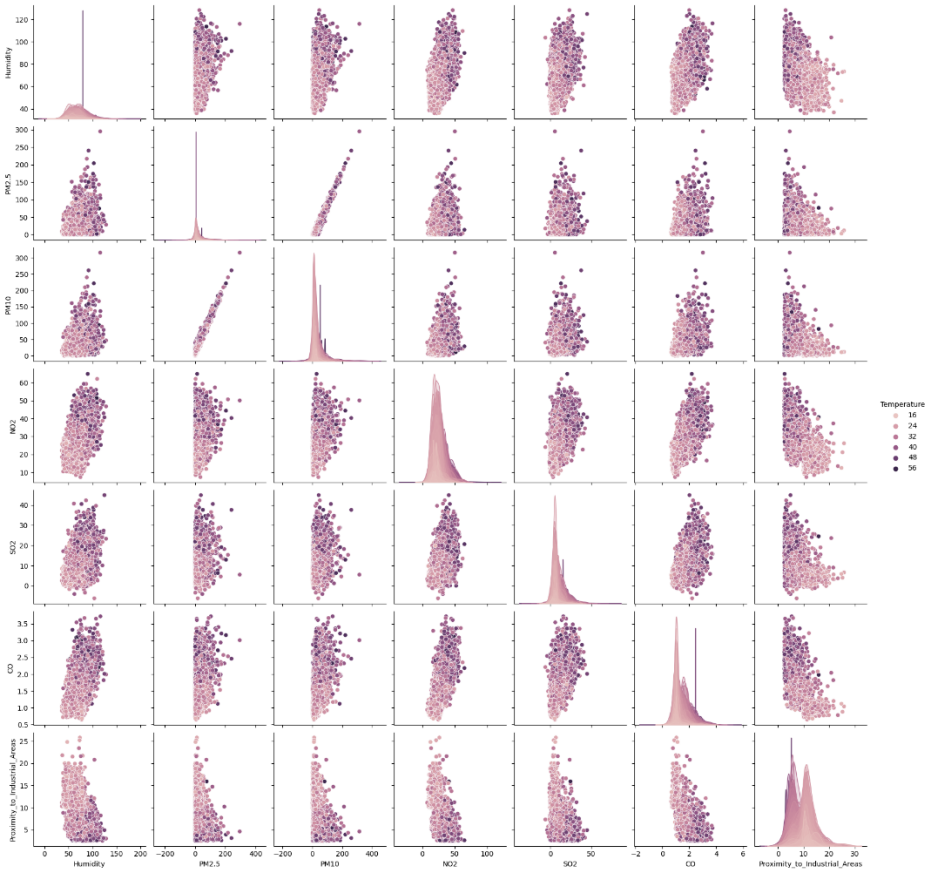

Figure 3. Scatter Plots after Normalisation

4. `metrics.classification_report` and `metrics.confusion_matrix`: For model evaluation.

Performance Evaluation

The model's performance is evaluated using two key metrics [15]–[18]:

1. **Classification Report**: Provides precision, recall, F1-score, and accuracy for each class. These metrics are calculated as [19]:

a. **Precision:** The ratio of true positive predictions to the total positive prediction:

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

b. **Recall:** The ratio of true positive predictions to the total actual positives:

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

c. **F1-Score:** The harmonic mean of precision and recall:

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (4)$$

2. **Confusion Matrix**: Represents the model's classification results in a tabular format to visualize correct and incorrect predictions [10], [12], [20], [21].

An optimal Decision Tree model is expected to achieve high precision and recall values, demonstrating its effectiveness in classifying air quality levels accurately.

3. Results and Discussion

Results

The classification model developed using the Decision Tree algorithm exhibited strong performance in categorizing air quality levels. The overall accuracy of the model reached 93%, indicating its ability to reliably classify air quality across four categories: Good, Moderate, Poor, and Hazardous. From the classification report, precision, recall, and F1-score were high across all classes, with particularly notable performance for the "Good" air quality category, achieving perfect scores in precision, recall, and F1-score. Classes such as "Moderate" and "Poor" demonstrated slightly lower but still robust metrics, with F1-scores of 0.94 and 0.82, respectively. These results underscore the model's reliability and balance in handling diverse air quality levels. **Table 2** show detailed classification report:

Table 2. Classification Reports

Class	Precision	Recall	F1-Score
Good (0)	1	1	1
Moderate (1)	0.93	0.95	0.94
Poor (2)	0.83	0.81	0.82
Hazardous (3)	0.82	0.82	0.82

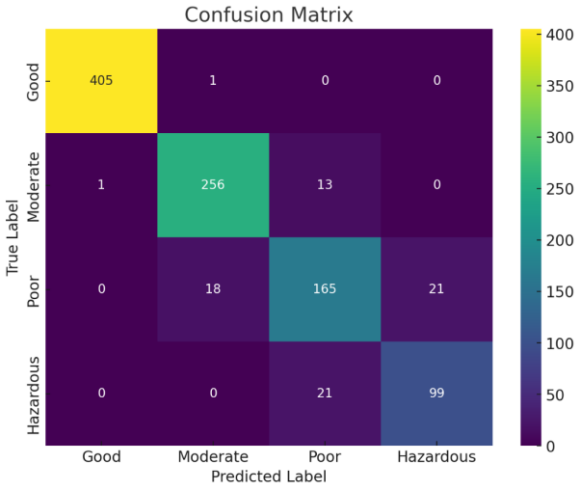

Figure 4. Confusion Matrix

Figure 4 visualization further confirms the model's performance. It illustrates the distribution of correct and incorrect predictions across all classes. The matrix shows minimal misclassifications, with the majority of predictions aligning with the true labels. For instance, there were only 13 misclassifications of "Moderate" air quality as "Poor," and only 21 instances of "Poor" air quality being misclassified as "Hazardous." These results emphasize the model's capability to distinguish between adjacent categories effectively.

The feature importance analysis revealed that PM2.5 concentration played a dominant role in predicting air quality levels, consistent with its widely recognized impact on pollution. Other significant features included population density and proximity to industrial areas, emphasizing the contribution of demographic and industrial factors to air quality. These findings are aligned with existing literature, which highlights the strong correlation between urban density, industrial emissions, and pollution levels.

Discussion

The results of this study align with previous research in the field of air quality analysis, where pollutants such as PM2.5 and PM10 have been identified as primary indicators of air quality degradation. The high classification performance, particularly for the "Good" and "Hazardous" categories, underscores the efficacy of the Decision Tree model in real-world applications. Furthermore, the model's interpretability makes it a valuable tool for policymakers and urban planners.

The findings also highlight the intricate relationships between environmental and demographic factors and air quality levels. High population density and proximity to industrial zones are consistently associated with poorer air quality, necessitating targeted interventions in these regions. Additionally, the model's ability to distinguish between different pollution levels offers potential for deploying early warning systems to protect sensitive populations.

In conclusion, the Decision Tree model provides a robust and interpretable framework for air quality classification. By combining high accuracy with actionable insights into feature importance, this study contributes to the development of effective strategies for pollution monitoring and mitigation. The inclusion of visualizations, such as the confusion matrix heatmap and scatter plots, enhances the interpretability of the results, making them accessible to a broad audience, including researchers, policymakers, and stakeholders.

4. Conclusion

This research successfully developed a Decision Tree-based classification model to predict air quality levels using environmental and demographic factors. The model achieved a high accuracy of 93%, with strong performance across all air quality categories. Key contributing factors, such as PM2.5 concentration, population density, and proximity to industrial zones, were identified as significant determinants of air quality. These findings provide actionable insights for policymakers and urban planners to design targeted interventions aimed at improving air quality and mitigating the adverse effects of pollution on public health and the environment. The visualizations, including the confusion matrix and feature importance rankings, further enhance the interpretability of the results, making them valuable for real-world applications.

Future research can build on these findings by incorporating additional environmental factors, such as wind speed, precipitation levels, and seasonal variations, to further refine the model's predictive capabilities. Integrating data from diverse geographical regions and exploring advanced ensemble methods, such as Random Forests or Gradient Boosted Trees, may also enhance classification performance. Moreover, the deployment of real-time monitoring systems based on such models can facilitate early warnings and enable more proactive air quality management strategies. These advancements would contribute significantly to the broader goal of sustainable urban development and public health protection.

References:

[1] D. Yassine, "Classification of Indoor CO2 Levels: Exploring the Impact of Humidity, Temperature, and Occupancy on Air Quality Using Machine Learning Model," *Proceedings of 2024 1st Edition of the Mediterranean Smart Cities Conference, MSCC 2024*. 2024, doi: 10.1109/MSCC62288.2024.10697053.

[2] E. Dossev, "Decision Trees for Event Signature Classification on Fiber Optic Cables in Quaternion Coordinates," *2022 European Conference on Optical Communication, ECOC 2022*. 2022.

[3] R. A. Raj, "Classification and Prediction of Incipient Faults in Transformer Oil by Supervised Machine Learning using Decision Tree," *2023 3rd International Conference on Artificial Intelligence and Signal Processing, AISP 2023*. 2023, doi: 10.1109/AISP57993.2023.10134566.

[4] I. Kilic, "Classification of Spyware from Network Packets with Decision Trees Using Recursive Feature Elimination (RFE)," *32nd IEEE Conference on Signal Processing and Communications Applications, SIU 2024 - Proceedings*. 2024, doi: 10.1109/SIU61531.2024.10600885.

[5] K. Kamyab-Hesari, "Machine learning for classification of cutaneous sebaceous neoplasms: implementing decision tree model using cytological and architectural features," *Diagn. Pathol.*, vol. 18, no. 1, 2023, doi: 10.1186/s13000-023-01378-w.

[6] Y. Chen, "Decision tree-based classification in coastal area integrating polarimetric SAR and optical data," *Data Technol. Appl.*, vol. 56, no. 3, pp. 342–357, 2022, doi: 10.1108/DTA-08-2019-0149.

[7] M. Aqib, "Classification of Edge Applications using Decision Tree, K-NN, & SVM Classifier," *2022 IEEE Students Conf. Eng. Syst. SCES 2022*, 2022, doi: 10.1109/SCES55490.2022.9887690.

[8] S. D. Permai, "Multiclass Classification for Air Quality In Jakarta Using Support Vector Machine and Multi-Layer Perceptron Classifier," *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*. pp. 198–202, 2022, doi: 10.1109/AiDAS56890.2022.9918697.

[9] S. Rani, "Machine Learning-based Multiclass Classification Model for Effective Air Quality Prediction," *2023 IEEE IAS Global Conference on Emerging Technologies, GlobConET 2023*. 2023, doi: 10.1109/GlobConET56651.2023.10149947.

[10] U. Zaky, A. Naswin, S. Sumiyatun, and ..., "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i3.113.

[11] D. Widyawati, A. Faradibah, and ..., "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.76.

[12] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data ...*, 2024, doi: 10.56705/ijodas.v5i1.126.

[13] M. N. Hasan, "Fetal Brain Planes Classification Using Deep Ensemble Transfer Learning from U-Net Segmented Fetal Neurosonography Images," *Int. J. Image, Graph. Signal Process.*, vol. 16, no. 4, pp. 74–86, 2024, doi: 10.5815/ijigsp.2024.04.06.

[14] T. T. Fousiya, "Diabetic Retinopathy Classification Based on Segmented Retinal Vasculature of Fundus Images Using Attention U-NET," *INDICON 2022 - 2022 IEEE 19th India Council International Conference*. 2022, doi: 10.1109/INDICON56171.2022.10039734.

[15] R. Rohan, "Classification of cardiac arrhythmia diseases from obstructive sleep apnea signals using decision tree classifier," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 12, pp. 248–264, 2020.

[16] D. R. Nemade, "Diabetes prediction using BPSO and decision tree classifier," *2nd Int. Conf. Data, Eng. Appl. IDEA 2020*, 2020, doi: 10.1109/IDEA49133.2020.9170744.

[17] I. A. P. Banlawe, "Decision Tree Learning Algorithm and Naïve Bayes Classifier Algorithm Comparative Classification for Mango Pulp Weevil Mating Activity," *2021 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2021 - Proc.*, pp. 317–322, 2021, doi: 10.1109/I2CACIS52118.2021.9495863.

[18] J. A. D. de Jesus Ferreira, "Decision tree classifiers for unmanned aircraft configuration selection," *Aircr. Eng. Aerosp. Technol.*, vol. 93, no. 6, pp. 1122–1132, 2021, doi: 10.1108/AEAT-03-2021-0074.

[19] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023, doi: 10.56705/ijaimi.v1i1.83.

[20] I. P. A. Pratama, E. S. J. Atmadji, and ..., "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indones. J. ...*, 2024, doi: 10.56705/ijodas.v5i1.124.

[21] S. Hidayat, H. M. T. Ramadhan, and ..., “Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification,” *Indones. J. ...*, 2023, doi: 10.56705/ijodas.v4i2.70.