International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i2.322

Research Article

Optimizing Air Quality Index Classification Using Multiple Machine Learning Models and Oversampling Techniques

Nuwairy El Furqany [1,*]
[1] Syiah Kuala University, Banda Aceh, Indonesia, nuwairy@gmail.com

Correspondence should be addressed to Nuwairy El Furqany; nuwairy@gmail.com

Received 04 September 2025; Revised 10 September 2024; Accepted 25 October 2024; Published 30 November 2025

Abstract:
Air quality significantly affects public health, environmental stability, and ecosystem balance. Accurate classification of the Air Quality Index (AQI) is critical for effective monitoring and management. Previous studies often relied on a single machine learning algorithm, which limited classification performance, particularly under class imbalance conditions. This study evaluates multiple machine learning algorithms for AQI classification, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, and Naïve Bayes. A random oversampling technique was applied to address the imbalance among AQI categories. The dataset consists of secondary data on pollutant concentrations (PM_{10}, SO_2, CO, O_3, NO_2) and AQI categories collected from five monitoring stations between 2010 and 2023. Model performance was assessed using accuracy, precision, recall, and F1-score. Before applying oversampling, the Random Forest model achieved an accuracy of 97.68%. After applying random oversampling, performance improved to 99.60%, with consistently high precision, recall, and F1-scores across classes. Feature importance analysis revealed that ozone (O_3) was the most influential pollutant, contributing 67.14% to model decision-making. The results demonstrate that combining random oversampling with ensemble-based machine learning substantially enhances AQI classification performance. This approach offers a robust and scalable framework for future air quality monitoring and environmental data analysis applications.

Keywords: Air Quality Index, Machine Learning, Classification, Random Oversampling, Random Forest.

Dataset link: https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021/

1. Introduction:

Air pollution is one of the most pressing environmental challenges of the modern era, with direct consequences for human health, environmental integrity, and ecosystem sustainability. Long-term exposure to polluted air has been linked to an increased risk of respiratory diseases, cardiovascular disorders, and premature mortality [1]. The decline in air quality is primarily driven by anthropogenic activities such as motor vehicle emissions, industrial processes, fossil fuel combustion, and land or forest fires. Even small-scale sources, such as household smoke from cigarettes, contribute to atmospheric contamination [2]. According to the 2024 IQAir report, Indonesia ranks among the 26 countries with the highest levels of air pollution worldwide, with the capital city Jakarta experiencing particularly severe conditions caused by pollutants including PM_{10}, NO_2, CO, SO_2, and O_3. These pollutants pose both acute and chronic health risks, especially for vulnerable populations such as children and the elderly [3].

Accurate assessment of the Air Quality Index (AQI) is essential for effective environmental monitoring and informed public health decision-making. In recent years, machine learning (ML) has emerged as a powerful tool for classifying AQI based on pollutant concentration data. Nevertheless, much of the existing research has relied on single-algorithm approaches, which limits performance optimization and reduces opportunities for comparative evaluation across different methods. For instance, Ramadhan and Triayudi [4] applied the C4.5 and Naïve Bayes algorithms to classify Jakarta's air quality based on the Air Pollutant Standard Index (ISPU), finding that the C4.5 algorithm outperformed Naïve Bayes with an average accuracy of 95%. Similarly, Razan et al. [5] employed the K-Means clustering algorithm to analyze air quality patterns in Jakarta, successfully identifying pollutant distribution trends and temporal variations that could guide targeted environmental management strategies. While these findings contribute valuable insights, they overlook a critical challenge in AQI classification: the frequent imbalance in datasets, which can bias model predictions toward majority classes and impair detection performance for underrepresented categories.

This study addresses these limitations by conducting a comprehensive comparison of multiple supervised machine learning algorithms Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, and Naïve Bayes for AQI classification. The research specifically incorporates the random oversampling technique to mitigate class imbalance and evaluates model performance using accuracy, precision, recall, and F1-score. Additionally, the study examines feature importance to identify the most influential pollutants in the classification process, thereby offering deeper insights into the dominant factors affecting urban air quality. The scope of this research is limited to AQI classification based on five pollutant concentrations PM_{10}, SO_2, CO, O_3, and NO_2 collected from five monitoring stations in Jakarta between 2010 and 2023. Meteorological variables such as temperature, humidity, and wind speed are excluded from the analysis. The contributions of this study are threefold: (1) providing a comparative evaluation of multiple machine learning algorithms for AQI classification, (2) assessing the effectiveness of random oversampling in improving classification performance under class imbalance, and (3) identifying the most influential pollutants contributing to AQI categorization.

The novelty of this study lies in its integrated framework that combines multiple supervised machine learning algorithms with random oversampling and feature importance analysis to enhance both classification performance and interpretability. Unlike previous studies that employed single-algorithm approaches without addressing class imbalance, this research provides a more balanced and explainable methodology for AQI classification. The proposed framework not only improves predictive accuracy under imbalanced data conditions but also highlights the relative importance of pollutants influencing AQI, thereby offering new insights for data-driven environmental management and air quality policy formulation.

2. Method:

The general research design of this study is illustrated in **Figure 1**, which presents the sequential workflow adopted to classify the Air Quality Index (AQI) using multiple machine learning algorithms. The research process consists of six main stages: (1) data preprocessing, (2) training and testing data processing, (3) data normalization and random oversampling, (4) multimodel machine learning implementation, (5) model performance comparison, and (6) feature

importance identification followed by conclusions and recommendations. Each stage is designed to ensure the reliability, fairness, and interpretability of the classification results through systematic evaluation.

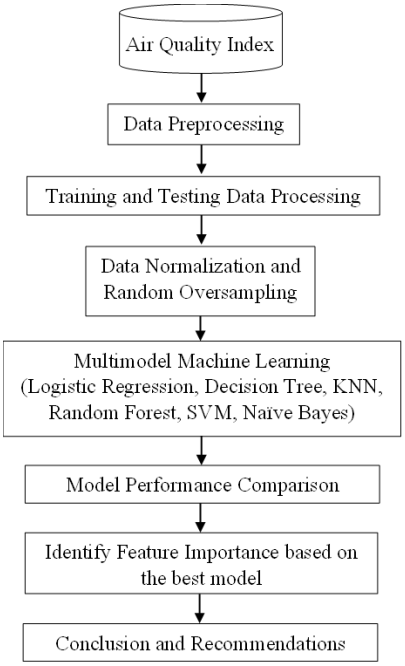

Figure 1. General Research Design

The study began with a data preprocessing stage that involved cleaning the dataset to address missing values, duplicates, and outliers. The dataset contained concentrations of PM_{10}, SO_2, CO, O_3, and NO_2 obtained from five air quality monitoring stations in DKI Jakarta between 2010 and 2023, with Air Quality Index (AQI) categories serving as the target variable. The data was then split into training and testing sets using an 80:20 ratio. All numerical variables were normalized using the Min–Max Scaler to ensure uniform feature scaling, and a random oversampling technique was applied to balance the class distribution and reduce model bias toward the majority class [6].

Subsequently, six supervised machine learning algorithms were implemented: Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, and Naïve Bayes. These algorithms were selected to represent both linear and nonlinear learning paradigms commonly used in environmental and air quality prediction research. Logistic Regression was chosen as a baseline linear model that provides interpretability and simplicity for assessing linear relationships between pollutants and AQI levels [7]. Decision Tree and Random Forest were included as tree-based ensemble models capable of capturing complex nonlinear interactions and threshold effects between pollutants [8]. K-Nearest Neighbors was applied as a distance-based algorithm sensitive to local variations in pollutant concentrations, while Support Vector Machine was employed for its strong generalization capability in high-dimensional, nonlinear classification tasks [9]. Finally, Naïve Bayes was used as a probabilistic baseline model, offering computational efficiency and robustness in handling mixed data distributions [10].

Model performance was evaluated using accuracy, precision, recall, and F1-score, both before and after the application of random oversampling, to assess the effect of class balancing on predictive performance. The best-performing model was further analyzed using a feature importance method to identify the relative contribution of each

pollutant to AQI classification. The findings from these stages served as the basis for drawing conclusions and providing practical insights for developing more accurate and interpretable air quality monitoring systems.

Data Collection:

The dataset used in this study consists of Air Quality Index (AQI) measurements for the Province of DKI Jakarta collected over a 13-year period (2010–2023). The data was obtained from the Kaggle open-data platform, which compiles verified air quality monitoring records from the Environmental Agency of DKI Jakarta. Observations were taken from five official Air Quality Monitoring Stations (SPKU) strategically distributed across the city to represent diverse urban characteristics and emission sources: Bundaran HI (Central Jakarta), Kelapa Gading (North Jakarta), Jagakarsa (South Jakarta), Lubang Buaya (East Jakarta), and Kebon Jeruk (West Jakarta). The dataset includes a total of 4,444 observations, each containing measurements of five major air pollutants PM_{10}, SO_2, CO, O_3, and NO_2 along with their corresponding AQI categories. Each record is classified into one of five AQI levels: Good, Moderate, Unhealthy, Very Unhealthy, and Hazardous. These pollutant indicators were selected because they represent the primary components used in global and national air quality assessment frameworks.

Table 1 summarizes the characteristics of each pollutant, including its unit of measurement, primary emission source, and typical health impacts.

Table 1. Feature Descriptions

Pollutant	Unit	Description	Health Impact
PM_{10}	$\mu g/m^3$	Fine particles from sources such as construction, industry, and natural dust	Respiratory problems
SO_2	ppb	Colorless gas from fuel combustion, industrial processes, and volcanic activity	Respiratory irritation
CO	ppm	Odorless gas from incomplete fuel combustion, mainly from vehicles	Reduced oxygen delivery, headache
O_3	ppb	Ground-level ozone formed from pollutants reacting with sunlight	Respiratory problems, plant damage
NO_2	ppb	Reddish gas mainly from vehicle emissions and combustion processes	Respiratory problems, lung impairment

As shown in Table 1, particulate matter (PM_{10}) and gaseous pollutants such as SO_2, CO, O_3, and NO_2 have direct implications for respiratory health and atmospheric reactivity. These parameters have been widely adopted as standard features in AQI modeling and prediction studies due to their strong correlation with pollution levels and health outcomes [10], [11], [12]. The inclusion of data from multiple stations across Jakarta ensures spatial representativeness, while the 13-year coverage provides sufficient temporal variation to support robust model training and validation. The distribution of air quality across the five monitoring stations in DKI Jakarta is illustrated in **Figure 2**.

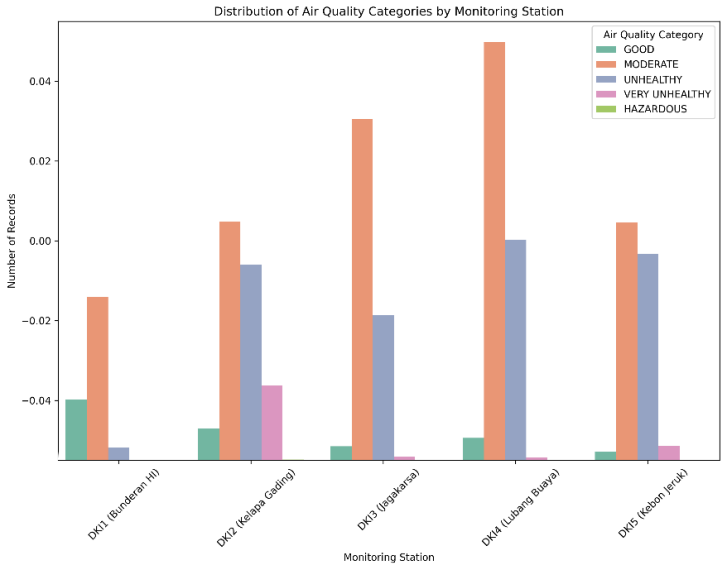

Figure 2. Air quality distribution in DKI Jakarta

The air quality distribution in DKI Jakarta, as shown in Figure 2, reveals significant variation among monitoring stations, illustrating differences in pollution levels that can occur across various locations within the city. The visualization results indicate that the “Moderate” category dominates the distribution across all monitoring stations, suggesting that most of the recorded air quality levels are generally acceptable for the majority of the population. However, they may begin to pose health risks to sensitive groups such as children, the elderly, or individuals with pre-existing respiratory conditions.

Furthermore, the DK4 (Lubang Buaya) station records the highest number of observations in the “Moderate” category compared to other stations, which may indicate that while the air quality in this area is relatively better, there is still significant pollution activity. On the other hand, the “Unhealthy” category is also notably prevalent at certain stations, such as DK2 (Kelapa Gading) and DK3 (Jagakarsa), indicating that these areas are frequently exposed to pollution levels that can be harmful to public health, especially for vulnerable populations. Variations in air quality distribution across stations may be influenced by multiple factors, including traffic density, industrial activities, and geographical conditions that affect pollutant dispersion [13]. These findings highlight the importance of more detailed and location-specific air quality monitoring to formulate more effective pollution control policies and to provide the public with accurate information regarding potential health risks associated with urban air pollution. The correlation matrix of pollutant concentrations used in AQI classification is presented in **Figure 3**.

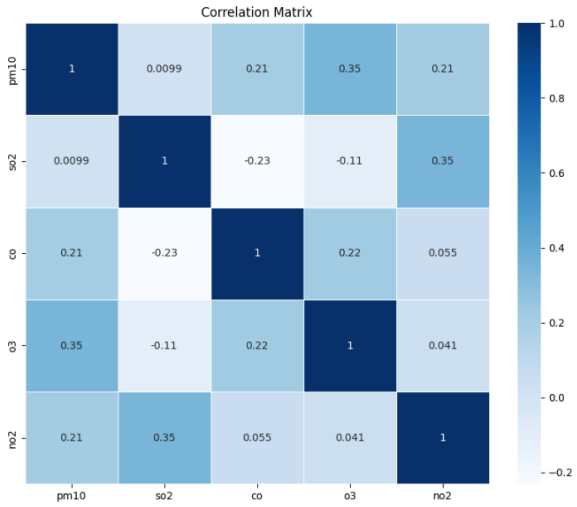

Figure 3. Correlation matrix of pollutant concentrations used in AQI classification

The correlation matrix in **Figure 3** illustrates the pairwise relationships among the five pollutants PM_{10}, SO_2, CO, O_3, and NO_2 used as features in the AQI classification model. The results indicate generally weak to moderate correlations between variables, suggesting that each pollutant provides distinct information for model learning. The highest positive correlation is observed between PM_{10} and O_3 (0.35), as well as between SO_2 and NO_2 (0.35), which may reflect similar emission sources or atmospheric interactions. A notable negative correlation is seen between SO_2 and CO (-0.23), possibly due to differences in their predominant sources or dispersion patterns. These findings imply that the inclusion of all five pollutants in the model is justified, as there is no evidence of strong multicollinearity that could adversely affect the classification process. Moreover, the modest correlations highlight the potential of machine learning models to capture non-linear relationships among variables that traditional linear models might overlook.

Data Preprocessing:

The dataset used in this study initially consisted of 4,626 samples. However, after a data cleaning process that removed entries containing missing values, the usable dataset was reduced to 4,444 samples. The dataset exhibited a significant imbalance in the distribution of air quality categories. To address this issue, the “Hazardous” and “Very Unhealthy” categories were merged into a single class labeled “Unhealthy” [14]. Following this merging process, the class distribution was as follows: “Unhealthy” with 1,660 samples, “Moderate” with 2,505 samples, and “Good” with 279 samples. In the initial stage of the research, the dataset was divided into two subsets: a training set and a testing set. A total of 90% of the samples (4,000 records) were allocated for training, while the remaining 10% (444 records) were used for testing. This 90:10 ratio was chosen to provide the model with a sufficiently large training set, enabling it to recognize patterns in the data effectively [15].

The preprocessing stage continued with the application of z-score normalization to standardize the scale across features, preventing features with larger value ranges from dominating the learning process. This step ensures that all features are treated equally by the machine learning algorithms, allowing the model to operate efficiently and accurately [16]. Subsequently, a random oversampling technique was applied to balance the number of samples across classes. This approach involved replicating samples from minority classes to increase their representation in the

training process. By employing this technique, the model becomes more sensitive to patterns within previously underrepresented classes, thereby improving fairness and accuracy in classification [17].

Class imbalance:

Class imbalance in the dataset, a common issue in many classification problems, can lead to biased models that tend to predict the majority class more frequently [18]. To address this issue, a random oversampling technique was applied, increasing the number of samples in minority classes so that the distribution across all classes became balanced [19], [20]. This approach aims to enhance the model's ability to learn patterns from all classes, including those that were previously underrepresented. The data distribution before and after applying the random oversampling technique is presented in **Table 2**. The table shows a significant adjustment in the number of samples for each air quality category, with minority classes increased to match the majority class size of 2,505 samples. This ensures that the machine learning models are not biased toward the majority class and can generalize more effectively for air quality classification tasks.

Table 2. Data Distribution Before and After Random Oversampling

Category	Actual Data	*Random Oversampling Data*
Good	279	2505
Moderate	2505	2505
Unhealthy	1461	2505
Very Unhealthy	198	2505
Hazardous	1	2505

In this study, random oversampling was applied to balance the distribution of samples across all categories by increasing the number of records in low-frequency categories to match the largest category size. The primary objective was to ensure proportional representation of each category so that the model would not be influenced by skewed class distributions. After balancing, the dataset was split into two main subsets: 80% for training and 20% for testing. This split allows the model to learn from most of the available data while ensuring evaluation is performed on unseen data, thereby providing a robust measure of generalization capability. By equalizing class sizes, the model is expected to be more sensitive to minority categories, ultimately improving classification accuracy and generalization performance.

Classification Algorithms and Performance Evaluation:

This study implements a variety of machine learning algorithms for air quality classification, namely Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Naïve Bayes. These algorithms were selected based on their capabilities and distinct characteristics in addressing classification problems, which are the primary focus of this research [21]. The chosen methods were considered suitable due to their ability to handle imbalanced data, ease of result interpretation, and strong generalization performance [22]. The dataset used in this research comprises Air Quality Index (AQI) measurements from the Province of DKI Jakarta. The primary objective is to analyze and classify air quality levels based on several relevant parameters, such as the concentrations of O_3, PM_{10}, SO_2, NO_2, and CO, which are known to affect public health.

Model performance evaluation was conducted using a confusion matrix, which provides a comparison between the model's predicted results and the actual conditions from the dataset. The confusion matrix enables an assessment of the model's ability to correctly classify data into the respective air quality categories [23]. Four primary evaluation metrics were employed accuracy, precision, recall, and F1-score to provide a comprehensive overview of the model's reliability in predicting air quality categories. The formulas for calculating each of these metrics are presented in Equations (1)-(4) and are used to evaluate the performance of each classification model [24].

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

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

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

$$F_1 - score = 2\ x\ \left(\frac{Precision\ x\ Recall}{Precision+Recall}\right)\ x\ 100\% \quad (4)$$

3. Result and Discussion:

Result

After applying the data balancing stage using random oversampling, six machine learning algorithms Logistic Regression, Decision Tree, KNN, Random Forest, SVM, and Naïve Bayes were implemented to analyze and classify the AQI data. Model performance was assessed using four evaluation metrics: accuracy, precision, recall, and F1-score, calculated from the confusion matrix. Table 3 presents the comparison of classification performance before and after applying random oversampling.

Table 3. Model Performance Before and After Random Oversampling

Algorithm	Before Oversampling				After Oversampling			
	Acc	**Prec**	**Recall**	**F1-score**	**Acc**	**Prec**	**Recall**	**F1-score**
Logistic Regression	85,46	85,39	85,46	86,27	88,34	88,08	88,34	88,17
Decision Tree	96,63	96,64	96,63	96,63	99,16	99,17	99,16	99,16
KNN	89,96	89,97	89,96	89,93	93,57	93,55	93,57	93,44
Random Forest	**97,68**	**97,68**	**97,68**	**97,68**	**99,60**	**99,60**	**99,60**	**99,60**
SVM	93,03	93,01	93,03	92,96	94,37	94,31	94,37	94,31
Naïve Bayes	87,86	87,96	87,86	87,84	89,82	89,59	89,82	89,66

The results show that before applying random oversampling, the Random Forest algorithm achieved the highest accuracy at 97.68%, followed by Decision Tree at 96.63%. Logistic Regression, Naïve Bayes, and SVM showed lower accuracy levels, indicating their sensitivity to class imbalance. After applying random oversampling, all algorithms improved in performance. Random Forest remained the top performer, with accuracy, precision, recall, and F1-score all increasing to 99.60%. Decision Tree followed closely with an accuracy of 99.16%. Other algorithms also improved substantially Logistic Regression reached 88.34%, KNN improved to 93.57%, and SVM rose to 94.37%. Naïve Bayes increased from 87.86% to 89.82%.

The improvement in model performance after oversampling demonstrates the effectiveness of data balancing in preventing bias toward majority classes. This finding is consistent with the results of [9], which also reported that ensemble-based algorithms such as Random Forest outperform single classifiers when trained on balanced environmental datasets. Similarly, [10] found that Random Forest yields superior predictive accuracy in urban AQI forecasting tasks due to its capacity to handle nonlinear pollutant interactions and noise within long-term datasets. Compared with these studies, the current research advances prior work by explicitly addressing data imbalance and conducting a broader comparative analysis across six machine learning algorithms using consistent preprocessing and evaluation procedures.

From a practical standpoint, achieving a classification accuracy of 99.6% indicates that the Random Forest model can correctly categorize nearly all air quality observations across five AQI levels. In real-world monitoring systems, this level of performance translates to high reliability in automated AQI reporting and early-warning systems, reducing misclassification risk in daily environmental updates and enabling timely public health responses. Moreover, high recall and F1-scores indicate that the model effectively detects minority AQI categories, which are often the most critical for environmental alerts, such as Unhealthy or Hazardous conditions. The comparison of machine learning model performance after applying random oversampling is illustrated in **Figure 4**.

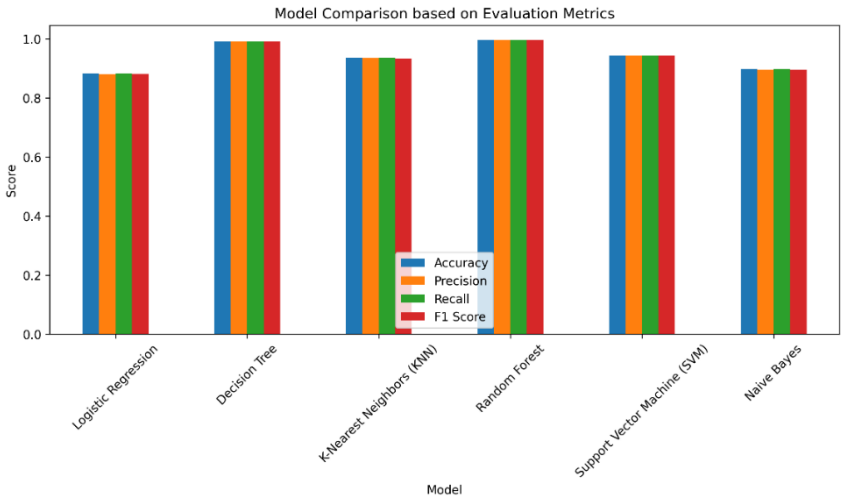

Figure 4. Comparison of machine learning models after applying random oversampling

The comparison in **Figure 4** visually illustrates the performance of the six classification algorithms after applying random oversampling. It is evident that Random Forest and Decision Tree achieved near-perfect scores across all evaluation metrics, indicating balanced performance between precision and recall as well as consistent classification accuracy. The confusion matrix of the best-performing model is presented in **Figure 5**.

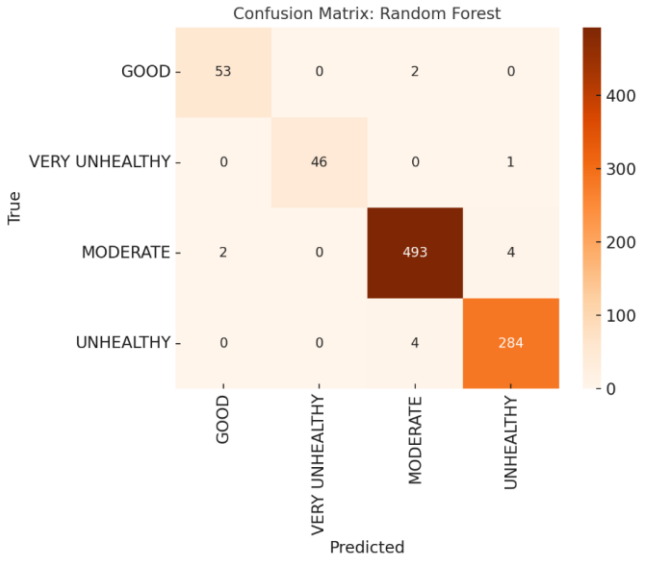

Figure 5. Confusion Matrix with the Best Model

As shown in **Figure 5**, the Random Forest model demonstrates exceptionally high classification accuracy across all AQI levels. The majority of observations fall along the diagonal, indicating strong predictive reliability. Specifically, the model correctly classified 53 samples in the "Good" category, 46 in "Very Unhealthy", 493 in "Moderate", and 284 in "Unhealthy." Only a few misclassifications occurred mainly between adjacent categories such as Moderate and Unhealthy which is expected due to overlapping pollutant concentration thresholds in real-world air quality data. This confusion matrix confirms that the Random Forest model can accurately differentiate between multiple AQI levels, maintaining strong discrimination even under conditions of data imbalance. The minimal number of false predictions highlights the model's robustness and ability to generalize from the training data. In practical terms, this performance implies that the Random Forest classifier can be confidently integrated into automated AQI monitoring systems to support reliable environmental reporting and early warning mechanisms for public health protection. The feature importance analysis derived from the Random Forest model is illustrated in **Figure 6**.

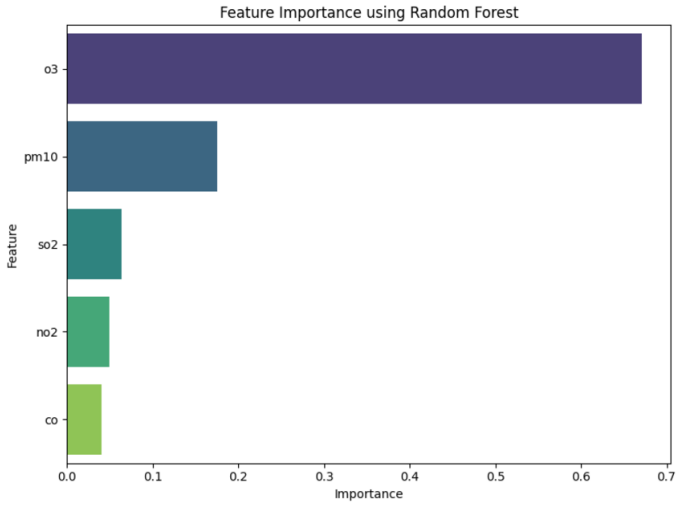

Figure 6. Feature importance analysis from the Random Forest model for AQI classification

The feature importance analysis in **Figure 6** shows that O_3 (ozone) is the most influential feature, contributing 67.14% to the model's decision process. This highlights the dominant role of ozone in determining air quality status in DKI Jakarta. Tropospheric ozone is a secondary pollutant formed through photochemical reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under sunlight, a process that is typically intensified in densely populated urban areas with high vehicular traffic and industrial activities. The second most important feature is PM_{10} (17.49%), which serves as a primary indicator of air quality in metropolitan environments. PM_{10} arises from motor vehicle emissions, industrial combustion, construction dust, and natural resuspension of road particles. Although other pollutants such as SO_2 (6.39%), NO_2 (4.99%), and CO (3.99%) contribute less numerically, their presence remains significant due to their established health effects and their role as precursors in atmospheric chemical reactions that contribute to secondary pollutant formation.

These findings confirm the interpretability advantage of ensemble learning models such as Random Forest, which not only provide strong predictive performance but also enable pollutant-level insight into model behavior. Consistent with previous studies [12], ozone and particulate matter remain the dominant pollutants shaping air quality variations in urban tropical climates. Consequently, air quality management in DKI Jakarta should prioritize efforts to reduce ozone and PM_{10} concentrations through integrated emission control policies, stricter vehicular standards, industrial process optimization, expansion of urban green spaces, and the promotion of sustainable public transportation systems.

Discussion

The results confirm that class imbalance in the dataset significantly affects model performance and predictive reliability. Before applying oversampling, models such as Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) exhibited bias toward the majority AQI classes, resulting in reduced accuracy and recall for minority categories. The application of random oversampling successfully equalized class representation, enabling all models to learn proportionally from each AQI category. Consequently, overall model performance improved substantially, particularly for algorithms previously sensitive to imbalance. The remarkable improvement observed in Random Forest (accuracy increasing from 97.68% to 99.60%) and Decision Tree (from 96.63% to 99.16%) highlights the effectiveness of tree-based ensemble methods in leveraging balanced datasets. These algorithms benefit from random feature selection and aggregation across multiple trees, allowing them to capture both linear and nonlinear pollutant relationships efficiently. Additionally, the performance gains seen in simpler models such as Logistic Regression and Naïve Bayes suggest that data balancing enhances generalization across different model complexities. This reinforces the importance of addressing class imbalance as a fundamental preprocessing step in environmental data modeling, rather than relying solely on algorithmic sophistication.

The feature importance analysis provides further insight into the environmental dynamics underlying AQI classification. Ozone (O_3) emerged as the most dominant factor, contributing 67.14% to model predictions. Tropospheric ozone is a secondary pollutant generated through photochemical reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under sunlight, a process amplified in tropical urban environments with high vehicular density and industrial emissions. Elevated ozone levels not only degrade air quality but also pose serious respiratory health risks, exacerbate asthma symptoms, and cause damage to vegetation and ecosystems. The second most influential pollutant, PM_{10}, accounted for 17.49% of the model's predictive importance. This finding

aligns with previous research identifying particulate matter as a key pollutant in metropolitan regions, primarily originating from transportation, industrial combustion, construction dust, and soil resuspension [12]. Although SO_2 (6.39%), NO_2 (4.99%), and CO (3.99%) exhibited smaller contributions, their environmental and health implications remain critical. SO_2 is associated with respiratory irritation and the formation of acid rain, NO_2 contributes to ozone and secondary particulate matter formation, and CO interferes with oxygen transport in the bloodstream, posing direct risks to cardiovascular health. Collectively, these pollutants interact synergistically within the atmospheric system, amplifying overall air quality deterioration.

From a policy perspective, these findings underscore the need for targeted interventions aimed at reducing ozone and PM_{10} concentrations in urban centers such as Jakarta. Recommended strategies include the implementation of stricter emission standards, industrial process optimization, and the adoption of cleaner production technologies. Expanding green infrastructure (e.g., urban trees, green corridors) can further mitigate air pollution through natural filtration, while promoting sustainable public transportation systems can substantially reduce vehicular emissions. The combination of a highly accurate predictive model and an interpretable feature importance framework establishes a robust foundation for data-driven air quality management. By integrating predictive analytics with environmental insight, policymakers can design adaptive, evidence-based interventions that simultaneously address pollution sources and safeguard public health. The study's results thus demonstrate the dual advantage of ensemble learning: analytical precision and actionable interpretability, both essential for modern urban sustainability planning.

4. Conclusion

This study compared the performance of six supervised machine learning algorithms Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Naïve Bayes for Air Quality Index (AQI) classification in DKI Jakarta. Before applying random oversampling, the Random Forest algorithm achieved the highest accuracy (97.68%), followed by Decision Tree (96.63%) and SVM (93.03%). Simpler algorithms such as Logistic Regression (85.46%), Naïve Bayes (87.86%), and KNN (89.96%) showed lower performance, reflecting their tendency to favor the majority class. After applying oversampling, all models improved substantially, with Random Forest attaining the highest and most consistent metrics (accuracy, precision, recall, and F1-score all at 99.60%). This confirms that data balancing enhances model generalization and fairness by allowing each algorithm to learn from all AQI categories proportionally.

The feature importance analysis identified O_3 (ozone) as the most influential pollutant (67.14%), followed by PM_{10} (17.49%), SO_2 (6.39%), NO_2 (4.99%), and CO (3.99%). These findings emphasize the need for continuous monitoring of ozone and particulate matter concentrations in urban environments, as they play a critical role in determining overall air quality. Despite the promising results, this study has several limitations. The dataset was limited to five monitoring stations within DKI Jakarta, which may constrain the spatial generalization of the model. Additionally, meteorological factors such as temperature, humidity, and wind speed known to influence pollutant dispersion were not included.

Future research should expand the dataset to cover multiple provinces or cities across Indonesia, allowing for cross-regional validation and assessment of model robustness under varying pollution profiles. Incorporating spatial–temporal modeling approaches (e.g., LSTM, CNN–LSTM hybrids) and advanced data-balancing techniques (e.g.,

SMOTEENN, ADASYN) could further improve predictive performance. Evaluating the model under real-time or streaming data conditions may also provide insights into its applicability for live AQI forecasting. From a policy perspective, the findings provide a strong data-driven foundation for targeted emission control strategies, particularly focusing on reducing ozone precursors and particulate matter emissions through stricter industrial standards, cleaner fuel adoption, and green transportation initiatives. Overall, this study demonstrates the combined benefits of data balancing and ensemble learning for building reliable, interpretable, and scalable machine learning frameworks in air quality management.

References:

[1] F. Islam, S. K. Nukala, P. Shrestha, T. Badgery-Parker, and F. Foo, "Air pollution and cardiovascular disease: A systematic review of the effects of air pollution, including bushfire smoke, on cardiovascular disease," *American Heart Journal Plus: Cardiology Research and Practice*, vol. 54, p. 100546, 2025, doi: 10.1016/j.ahjo.2025.100546.

[2] J. Guo, G. Chai, X. Song, H. Xu, Z. Li, X. Feng, and K. Yang, "Long-term exposure to particulate matter on cardiovascular and respiratory diseases in low- and middle-income countries: A systematic review and meta-analysis," *Frontiers in Public Health*, vol. 11, p. 1134341, 2023, https://doi.org/10.3389/fpubh.2023.1134341.

[3] IQAir, "Kualitas udara di Indonesia," Nov. 15, 2024. [Online]. Available: https://www.iqair.com/id/indonesia.

[4] D. P. Ramadhan and A. Triayudi, "Jakarta air quality classification based on air pollutant standard index using C4.5 and Naïve Bayes algorithms," *Journal of Technology and Information Systems*, vol. 2, no. 4, 2024, https://doi.org/10.58905/saga.v2i4.395.

[5] M. A. F. Razan, N. J. Alifah, Q. A'yuni, M. Wati, and H. -, "Application of K-Means Clustering Algorithm for Air Quality Pattern Analysis in Jakarta," *JUTIKOMP*, vol. 8, no. 1, pp. 64–80, 2025, https://doi.org/10.34012/jutikomp.v8i1.7028.

[6] F. Liu and Y. Dai, "Product processing quality classification model for small-sample and imbalanced data environment," *Computational Intelligence and Neuroscience*, vol. 2022, p. 9024165, 2022, https://doi.org/10.1155/2022/9024165.

[7] C. Shi, Y. Wang, Y. Wan and S. Wu, "Air Quality Prediction Based on Machine Learning," 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, 2022, pp. 1-5, https://doi.org/10.1109/MLKE55170.2022.00008.

[8] I. G. Iwan Sudipa, M. Habibi, E. S. Jullev Atmadji, and I. Arfiani, "Predictive Modeling of Air Quality Levels Using Decision Tree Classification: Insights from Environmental and Demographic Factors", ijodas, vol. 5, no. 3, pp. 251-258, Dec. 2024, https://doi.org/10.56705/ijodas.v5i3.201.

[9] M. Karmoude, B. Munhungewarwa, I. Chiraira, R. Mckenzie, J. Kong, B. Smith, G. Ayana, N. Njara, T. Mathaha, M. Kumar, and B. Mellado, "Machine learning for air quality prediction and data analysis: Review

on recent advancements, challenges, and outlooks," Science of The Total Environment, 2025, https://doi.org/10.1016/j.scitotenv.2025.180593.

[10] S. Ameer et al., "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," in IEEE Access, vol. 7, pp. 128325-128338, 2019, https://doi.org/10.1109/ACCESS.2019.2925082.

[11] Z. Chen, N. Liu, H. Tang, X. Gao, Y. Zhang, H. Kan, F. Deng, B. Zhao, X. Zeng, Y. Sun, H. Qian, W. Liu, J. Mo, X. Zheng, C. Huang, C. Sun, and Z. Zhao, "Health effects of exposure to sulfur dioxide, nitrogen dioxide, ozone, and carbon monoxide between 1980 and 2019: A systematic review and meta-analysis," *Indoor Air*, vol. 32, no. 11, p. e13170, 2022, https://doi.org/10.1111/ina.13170 .

[12] P. Vongelis, N. G. Koulouris, P. Bakakos, and N. Rovina, "Air pollution and effects of tropospheric ozone (O_3) on public health," *International Journal of Environmental Research and Public Health*, vol. 22, no. 5, p. 709, 2025, https://doi.org/10.3390/ijerph22050709.

[13] J. S. Ji, L. Liu, J. Zhang, et al., "NO_2 and $PM_{2.5}$ air pollution co-exposure and temperature effect modification on pre-mature mortality in advanced age: a longitudinal cohort study in China," Environmental Health, vol. 21, no. 97, 2022, https://doi.org/10.1186/s12940-022-00901-8.

[14] S. Sohrab, N. Csikós, and P. Szilassi, "Effect of geographical parameters on PM10 pollution in European landscapes: a machine learning algorithm-based analysis," Environmental Sciences Europe, vol. 36, no. 152, 2024, https://doi.org/10.1186/s12302-024-00972-z.

[15] F. Hamami and I. A. Dahlan, "Air quality classification in urban environment using machine learning approach," *IOP Conference Series: Earth and Environmental Science*, vol. 986, no. 1, p. 012004, 2022, https://doi.org/10.1088/1755-1315/986/1/012004.

[16] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, 2022, https://doi.org/10.1002/sam.11583.

[17] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109924, 2023, https://doi.org/10.1016/j.asoc.2022.109924.

[18] R. Idroes, et al., "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, p. 137, 2021, https://doi.org/10.15866/iremos.v14i2.20460.

[19] N. El Furqany, M. Subianto, and A. Rusyana, "Hybrid ensemble learning with SMOTEENN and soft voting for stunting risk prediction: A SHAP-based explainable approach," Journal of Applied Data Sciences, vol. 6, no. 4, pp. 2989–3004, Dec. 2025, https://doi.org/10.47738/jads.v6i4.829.

[20] W. Chen, K. Yang, Z. Yu, et al., "A survey on imbalanced learning: latest research, applications and future directions," Artificial Intelligence Review, vol. 57, p. 137, 2024, https://doi.org/10.1007/s10462-024-10759-6.

[21] Y. B. Wah, et al., "Machine learning and synthetic minority over-sampling techniques for imbalanced data: Improving machine failure prediction," *Computers, Materials & Continua*, vol. 75, no. 3, pp. 4821–4841, 2023, https://doi.org/10.32604/cmc.2023.034470.

[22] J. Han, et al., *Data mining: Concepts, models, methods, and algorithms*, 3rd ed. Elsevier; Morgan Kaufmann, 2012.

[23] V. Chang, J. Bailey, Q. A. Xu, T. Li, and X. Cao, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, vol. 35, no. 24, pp. 16157–16173, 2023, https://doi.org/10.1007/s00521-022-07049-z.

[24] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, 2022, https://doi.org/10.1007/s00521-022-07049-z.

[25] J. Kozak, B. Probierz, K. Kania, and P. Juszczuk, "Preference-driven classification measure," Entropy, vol. 24, no. 4, p. 531, 2022, https://doi.org/10.3390/e24040531.