Development of an Air Quality Classification System Using SMOTE-Based Random Forest and XAI Analysis

[1]Arip Kristiyanto, [2]Hirawati Lubis
[1]Information Systems, Universitas Pamulang, Serang
[2] Mathematics, Universitas Pamulang, Serang
[1]dosen10027@unpam.ac.id, [2]hirawati.lubis@unpam.ac.id

Abstract - South Tangerang City is a critical environmental issue that requires an accurate and transparent classification system. This study aims to develop an air quality classification model using a machine learning algorithm integrated with data balancing techniques and model interpretation methods. The methodology used includes pre-processing of Air Pollutant Standard Index (ISPU) data for the 2020–2022 period into three categories: Good, Moderate, and Unhealthy. The dataset used is 1096, Synthetic Minority Over-sampling Technique (SMOTE) is applied to handle class imbalance, and hyperparameter optimization is performed using GridSearchCV. The experimental results show that the Random Forest algorithm outperforms the baseline SVM and KNN models, achieving an accuracy of 0.81 and an F1-Score of 0.75 after SMOTE and tuning. Explainable AI (XAI) analysis using SHAP reveals that sulfur dioxide (SO_2) is the most dominant feature influencing model decisions, and it is spatially correlated with industrial activities and heavy transportation in the South Tangerang area. The final model was then deployed to the Hugging Face Spaces cloud platform via the Gradio interface to provide publicly accessible classification services. This study demonstrates that integrating Random Forests and SHAP produces a classification system that is not only highly performant but also scientifically transparent, supporting air pollution mitigation.

Keywords — Air Quality Classification, Random Forest, SMOTE, SHAP , Explainable AI

I. Introduction

As a global environmental issue, air pollution has a serious impact on human health and the balance of ecosystems. The World Health Organization states that long-term exposure to air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) can increase the risk of health problems and premature death [1][2]. In urban areas, increased transportation and industrial activity exacerbate air pollutant concentrations, necessitating an accurate and responsive air quality monitoring and classification system.

South Tangerang City, as a buffer zone for Jakarta, is experiencing rapid growth in the residential, transportation, and industrial sectors. High community mobility and dense traffic make the transportation sector a major contributor to air pollution in this region. Furthermore, the presence of manufacturing industries, such as paper processing and ceramics, and other sectors, also contributes to increased pollutant emissions. This condition is further exacerbated by cross-regional pollution from the Greater Jakarta metropolitan area [3].

With advances in computing technology, machine learning (ML) approaches have been widely used for air quality prediction and classification. Several previous studies have reported that ensemble-based algorithms, such as Random Forest, XGBoost, and LightGBM, demonstrate superior performance compared to conventional statistical methods. This is due to their ability to model non-linear relationships and manage feature complexity in the data [4]. A comprehensive study conducted by Rybarczyk confirmed that ensemble-based models tend to be stable on environmental data with high variability [5].

However, several challenges arise when applying machine learning to air quality data. First, the class distribution in air quality datasets is often imbalanced, with certain categories, such as "Good" or "Moderate," being more dominant than the "Unhealthy" category. This imbalanced data distribution can lead the model to overestimate the majority class, thereby reducing classification performance for the minority class. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) is used to address this imbalance. This method is widely used and has proven effective at improving model performance, particularly in classifying minority-class data [6][7].

Second, although Random Forest is known for its high accuracy and resistance to overfitting [8], it is often considered a black-box model due to its lack of interpretability. In the context of environmental decision support systems, model transparency is crucial for ensuring that prediction results are understandable to policymakers [9]. Therefore, Explainable Artificial Intelligence (XAI) approaches and model interpretation methods, such as SHAP (Shapley Additive exPlanations), are used to explain the contribution of each feature to the model's predictions [10][11]. This approach has been widely applied in the analysis of modern air quality prediction models [12] [13].

In addition to accuracy and interpretability, real-world system implementation is also an important factor in current research. Several recent studies focus not only on model development but

also on integrating web- or cloud-based systems to support real-time inference [14][15].

Although various previous studies have attempted to improve the accuracy of air quality classification using ensemble algorithms such as Random Forest, the main challenge remains the nature of air pollution datasets, which often suffer from class imbalance. Oversampling techniques such as SMOTE are commonly used to address this issue, but their effectiveness is limited in datasets with high feature overlap, such as urban ISPU parameters. Furthermore, model transparency is often overlooked; even high-accuracy models remain black boxes, making it difficult for policymakers to understand which pollutants most influence changes in local air quality. Therefore, this study addresses this gap by critically evaluating the impact of SMOTE on an optimized Random Forest model, integrating SHAP-based Explainable AI (XAI) to provide more scientific and accountable feature interpretation, and implementing the approach in a cloud architecture.

II. Research Method

This study used an experimental approach to evaluate classification accuracy using the Random Forest algorithm. The model optimization process used hyperparameter tuning, while data imbalance was addressed with SMOTE. Furthermore, a SHAP-based Explainable AI approach was used to increase model transparency.

The following are the research stages, from data collection to model deployment, as shown in Figure 1.

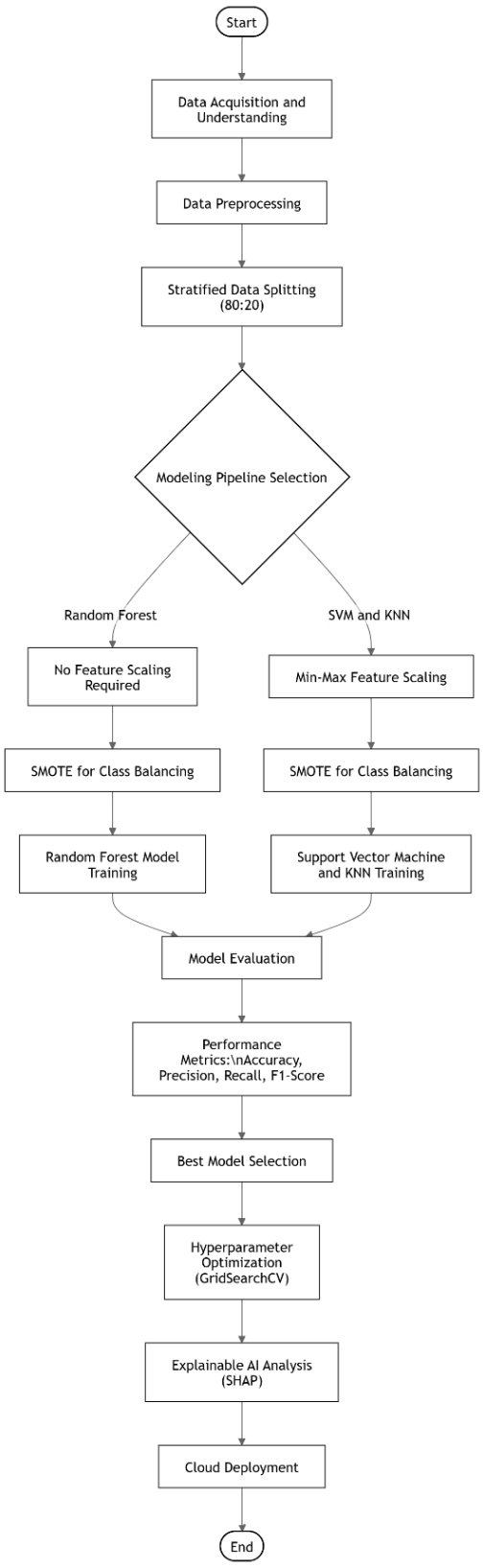


Figure 1. Research Stages

A. Data Understanding

The dataset used is the Air Quality Index data in South Tangerang City for the period 2020–2022[16]. The dataset shown in Figure 2 consists of 1,096 data points with six pollutant features: PM2.5, PM10, CO, SO_2, O_3, dan NO_2.

	Date	PM2.5	PM10	SO2	CO	O3	NO2	Max	Critical Component	Category
0	1/1/2020	45	30	2	69	19.0	0	69	CO	Moderate
1	1/2/2020	44	16	2	58	33.0	0	58	CO	Moderate
2	1/3/2020	43	12	2	46	18.0	0	46	CO	Good
3	1/4/2020	40	8	2	84	29.0	0	84	CO	Moderate
4	1/5/2020	38	8	3	50	0.0	0	50	CO	Good

Figure 2. Dataset

The target variables or classes are Good, Moderate, and Unhealthy. Figure 3 shows the distribution of data for each class.

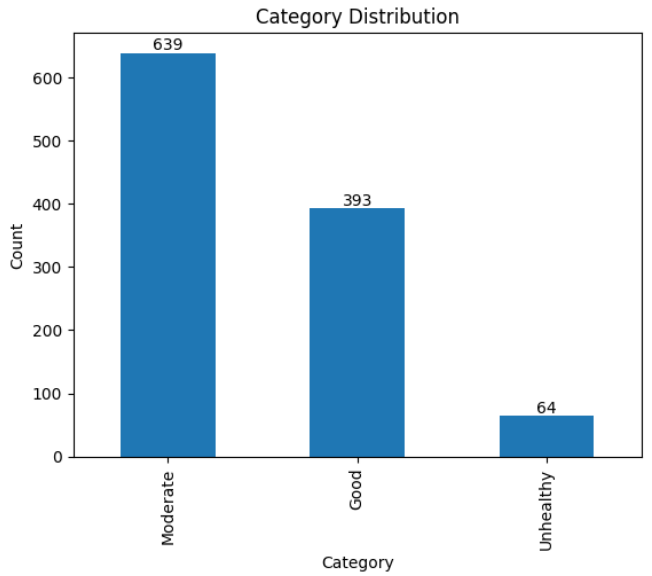


Figure 3. Class Distribution

B. Data Preprocessing

This stage improves the quality of the model output by performing several steps:

Removing irrelevant features, missing values, and duplicate data.

a. Missing data handling is performed using the dropna() function. The decision to remove rows containing blank values is based on initial analysis, which showed that the percentage of missing data is relatively small, at 5% of the original dataset of 1,096 samples. This method is considered safer than simple imputation techniques (such as mean or median imputation) to avoid artificial noise in the data distribution that can obscure natural correlations between pollutants.

b. Removing irrelevant features using the drop column function removes the Date, Max, and Critical Component features.

No duplicate data was found in the dataset. After cleaning, the final dataset consisted of 1,036 rows, with a class distribution of 579 Moderate, 393 Good, and 64 Unhealthy.

C. Train/Test Split

The preprocessed dataset was then divided into two parts: a training set (80%) and a testing set (20%). This stratified division was performed to ensure that the ratio of air quality labels in both sets remained representative of the original dataset.

D. Data Normalization

In this study, to identify the best model for deployment in the air quality classification system, a performance comparison was conducted between the proposed Random Forest model and two baseline algorithms: Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Since SVM and KNN are distance-based algorithms that are sensitive to variations in feature scale, a normalization step was performed using the Min-Max Scaling method specifically for these two algorithms. This normalization transforms all ISPU parameters into a range of values from 0 to 1 using the equation. The original data was retained for Random Forest.

E. SMOTE

This study integrates the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Given the significant class imbalance among the 'Good', 'Moderate', and 'Unhealthy' categories, SMOTE was applied to generate synthetic samples for the minority class. SMOTE was applied after the data split and to the training data [17][18].

F. Modeling

This study compared three types of machine learning algorithms: Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). All three were trained using historical air pollution data from South Tangerang from 2020 to 2022, then tested for their classification performance. The best model was then subjected to hypertuning and deployment.

G. Evaluation

Classification model performance was evaluated using several performance metrics commonly used in machine learning: accuracy, precision, recall, and F1 score [7].

a. Accuracy

Accuracy is a model performance metric, calculated by dividing the number of correct predictions by the total number of predictions generated by the classification model.

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

True Positive (TP) refers to correctly predicted positive data, True Negative (TN) refers to correctly predicted negative data, False Positive (FP) refers to incorrectly predicted negative data, and False Negative (FN) refers to incorrectly predicted positive data.

b. Precicion

Precision measures the ratio of accurate positive predictions to the total number of positive predictions generated by the model.

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

c. Recalls

Recall measures the ratio of actual positive predictions to the total number of actual positive cases in the dataset.

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

d. F1-Score

The harmonic mean of precision and recall offers a balance between these two criteria.

$$\boldsymbol{F1-Score} = \boldsymbol{2} \times \frac{\boldsymbol{Precicion \times Recall}}{\boldsymbol{Precicion+Recall}} \quad (4)$$

H. Hyperparameter

The optimization process aims to find the best parameter combination that minimizes generalization error in the Random Forest model. The initial parameter range for GridSearchCV was determined based on the characteristics of the ISPU dataset, which has quite dynamic feature variance but a limited number of samples (1,036 rows).

I. SHAP *Analysis*

Interpretability analysis was performed using SHAP to determine the contribution of features to predictions using SHAP interaction values and mean SHAP interaction values. SHAP interaction values measure the contribution of feature interactions to predictions at the individual level, enabling an in-depth analysis of the influence of feature combinations on each instance. The mean SHAP interaction value provides an overview of the relationships between features in the model.

$$Mean\ SHAPinteraction = \frac{1}{n}\sum_{i=1}^{n} SHAP^{i} \quad (5)$$

The mean SHAP value is the average SHAP value for each feature across the entire dataset, used to measure a feature's global importance in the model. Unlike SHAP values, which are local and describe a feature's contribution to a specific instance, the mean SHAP value provides a general overview of a feature's influence on the overall model prediction. The feature with the highest mean SHAP value is interpreted as the most dominant variable influencing the model output.

$$Mean\ SHAP_j = \frac{1}{n}\sum_{i-1}^{n} \left|\emptyset_j^{(i)}\right| \quad (6)$$

n is the total number of data points,

$\phi_j^{(i)}$ is the SHAP value of j^{th} feature at the i^{th}.

J. Deployment

The system was implemented in Python using the Gradio framework and deployed to the Hugging Face Spaces cloud platform to provide an interactive, publicly accessible web-based classification service [19]. This system still has operational limitations in terms of data acquisition. Currently, the model is not directly integrated with wireless sensor networks or Internet of Things (IoT) devices in the field, so the data update process still relies on manual input from the ISPU historical database.

III. Results and Discussion

A. Model Evaluation

Tabel 1 below presents the evaluation results before using SMOTE for each model, including Accuracy, Precision, Recall, and F1 Score.

Tabel 1. Model Comparison Evaluation

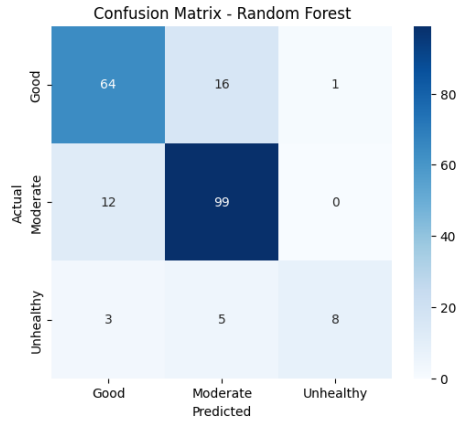
Model	Normalization	Accuracy	Precision	Recall	F1-Score
RF	No	0.82	0.84	0.73	0.77
KNN	Min-Max	0.79	0.72	0.66	0.68
SVM	Min-Max	0.72	0.68	0.57	0.59

An initial evaluation was conducted to establish a baseline performance standard by comparing three classification algorithms: Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), before applying data balancing and hyperparameter optimization techniques. Based on the test results presented in Table 1, Random Forest demonstrated superiority over the other models, achieving an accuracy of 0.82, a precision of 0.84, a recall of 0.73, and F1-score of 0.77. Random Forest's superiority in processing the ISPU dataset is influenced by its scale-invariant architecture and ensemble learning mechanism, which is inherently more robust to data noise and outliers than distance-based models or single hyperplanes.

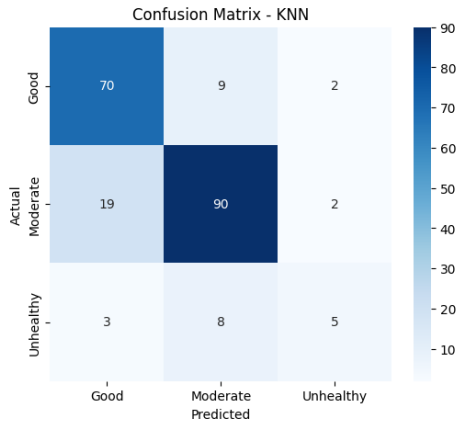
In contrast, the KNN model with Min-Max Scaling normalization achieved moderate performance, with an accuracy of 0.79 and an F1-score of 0.68. Although normalization was used to align the pollutant feature scales, KNN's performance was lower than RF's, indicating that the distance-based algorithm is highly sensitive to local data density and distribution, particularly in areas of class overlap that are common in air quality parameters. Meanwhile, SVM recorded the lowest performance with an accuracy of 0.72 and an F1-score of 0.59. The low efficacy of SVM at this stage is strongly suspected to be due to the use of default parameters, which are not yet able to optimally map the decision boundary in the complex feature space.

Overall, these results indicate that without addressing data imbalance and model optimization, Random Forest is the most stable and superior algorithm. This finding also emphasizes the importance of advanced preprocessing steps, such as data balancing and hyperparameter tuning, to improve model performance, especially in algorithms that are sensitive to data distribution, such as KNN and SVM.

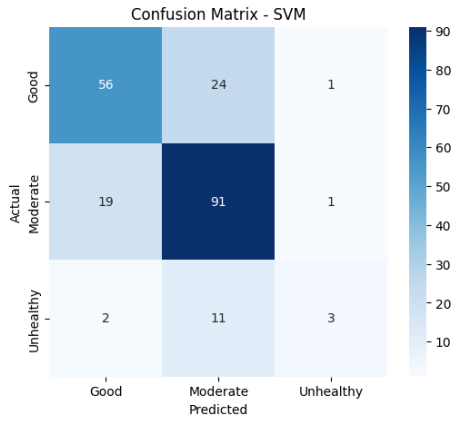
Based on the confusion matrix results, the three models showed different abilities in classifying air quality data into three classes: Good, Moderate, and Unhealthy.



(a) Random Forest



(b) KNN



(c)SVM

Figure 4. Comparison of Confusion Matrix (a) Random Forest, (b) KNN, (c) SVM

The Random Forest model demonstrated the best overall performance. In the Moderate class, the model correctly classified 99 data points, the highest score among the models. In the Good class, there were 64 correct predictions with a lower error than the SVM. Furthermore, performance in the Unhealthy class was also better than KNN and SVM, with 8 data points correctly classified. Random Forest's classification errors tended to be lower and more balanced across classes, indicating better generalization ability.

The KNN model performed quite well on the Good class, correctly classifying 70 data points, but still misclassified 9 data points to the Moderate class. In the Moderate class, the model correctly classified 90 data points but still misclassified 19 as the Good class. The main weakness of KNN was observed in the Unhealthy class, where only 5 data points were correctly classified, while the majority 8 data were misclassified as the Moderate class. This indicates that KNN is suboptimal at recognizing minority classes.

The SVM model improved on the Moderate class, correctly classifying 91 data points. However, performance in the Good class decreased compared to KNN, with only 56 correct predictions and significant errors in the Moderate class (24 data points).

In the Unhealthy class, SVM performance was relatively low, with only 3 data points correctly classified and the majority misclassified as the Moderate class (11 data points). This indicates that SVM tends to be biased toward the majority class (Moderate).

B. Impact of SMOTE Implementation

Based on Table 2, the class distribution before SMOTE implementation showed significant class imbalance. The Moderate class had the most data points (468), followed by the Good class (312), and the Unhealthy class was the minority class with only 48 data points.

To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was implemented. After SMOTE was implemented, the number of data points per class was balanced at 468 for each of the Good, Moderate, and Unhealthy classes.

Tabel 2. Class Distribution Before and After SMOTE

Label	Class	Before SMOTE	After SMOTE
0	Good	312	468
1	Moderate	468	468
2	Unhealthy	48	468

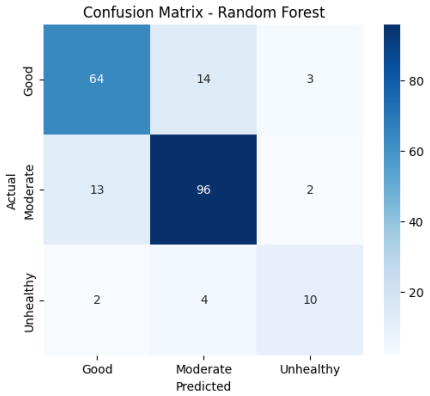
Tabel 3. below presents the evaluation results for each model after applying SMOTE, including Accuracy, Precision, Recall, and F1 Score.

Tabel 3. Model Comparison Evaluation After SMOTE

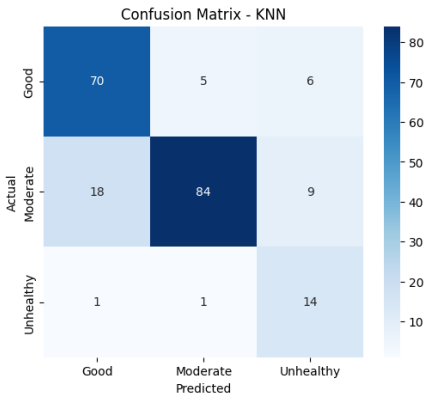
Model	Normalization	Accuracy	Precision	Recall	F1-Score
RF+S	No	0.82	0.77	0.76	0.77
KNN+S	Min-Max	0.80	0.73	0.83	0.76
SVM+S	Min-Max	0.70	0.64	0.76	0.67

Based on Table 3, an evaluation was conducted after applying SMOTE to address class imbalance. The results show that Random Forest achieved the best performance, with an accuracy of 0.82, precision of 0.77, recall of 0.76, and F1-score

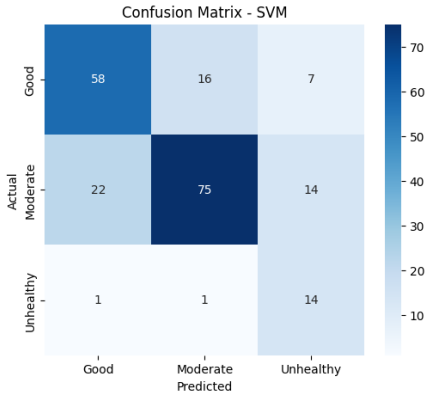
of 0.77. This indicates that Random Forest is able to maintain stable performance despite the data being oversampled.
The KNN model achieved an accuracy of 0.80, the highest recall (0.83), but lower precision (0.73). This indicates that KNN is quite good at capturing most of the positive data, but still produces inaccurate predictions (relatively high false positives).
Meanwhile, SVM performed worst, with an accuracy of 0.70, precision of 0.64, a recall of 0.76, and an F1-score of 0.67. Despite the good recall, the low precision value indicates that this model is less able to accurately distinguish between classes after the SMOTE process.
Based on the evaluation results, Random Forest was the most optimal model after applying SMOTE, both in terms of accuracy and the balance of evaluation metrics. However, the use of SMOTE should be carefully considered, as it does not always improve model performance, especially on datasets with overlapping feature distributions. Evaluation using the confusion matrix in Figure 4 reveals significant differences in the generalization ability of the three models across air quality categories.



(a) Random Forest



(b) KNN



(c) SVM

Figure 5. Perbandingan Confusion Matrix Setelah SMOTE (a) Random Forest, (b) KNN, (c) SVM

The Random Forest model still demonstrated the best overall performance. In the Moderate class, the model correctly classified 96 data points, the highest score among the models. In the Good class, there were 64 correct predictions, while the Unhealthy class had the lowest performance, with 10 data points correctly classified.
The KNN model performed quite well on the Good class, correctly classifying 70 data points and incurring 18 errors. In the Moderate class, the model correctly classified 84 data points, with 5 errors. In the Unhealthy class, the model improved over the previous model, correctly classifying 14 data points. The SVM model performed quite well on the Good class, correctly classifying 58 data points compared to 70 previously and incorrectly classifying 22. In the Moderate class, the model correctly classified 75 data points and misclassified 16. In the Unhealthy class, there was an improvement over the previous class, with 14 data points correctly classified and 1 misclassified.
Figure 6 further presents the trend in accuracy achieved by the three classification algorithms. This visualization compares model performance under baseline conditions (without SMOTE) with that after SMOTE integration.

Figure 6 shows that the Random Forest model performed better than K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), both before and after SMOTE. Before data balancing, Random Forest achieved the highest accuracy and F1-score, followed by KNN and SVM. After applying SMOTE, recall values increased significantly across all models, especially for KNN and SVM, indicating improved ability to detect minority classes. However, this increase in recall was accompanied by a decrease in precision, indicating an increase in false positives due to oversampling. Overall, Random Forest remained the most stable model, with consistent performance across all evaluation metrics, while KNN showed the greatest improvement after SMOTE. Experimental results showed that applying SMOTE did not significantly improve accuracy values for all tested models. This is because accuracy tends to be

biased towards the majority class, so changing the data distribution through oversampling does not directly improve this metric.

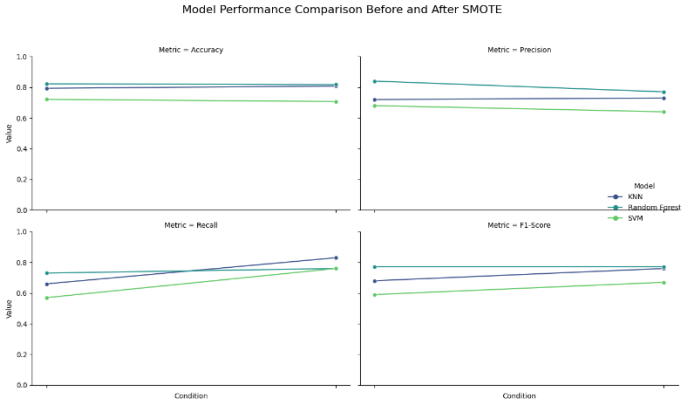


Figure 6. Performance Comparison Before and After SMOTE

C. Hyperparameter Tuning Results

After selecting the best model, the Random Forest, the next step was to optimize it to find the parameter combination that minimized generalization error. The initial parameter range for Grid Search CV was determined based on the characteristics of the ISPU dataset, which has quite dynamic feature variance but a limited sample size (1,036 rows). Table 4 below details the parameters tested in the tuning process.

Tabel 4. Tuning Process Parameters

Hyperparameter	Value/Range
n_estimators	[100, 200]
max_depth	[None, 10, 20]
min_samples_split	[2, 5]
min_samples_leaf	[1, 2]

Tabel 5. RF Model Evaluation After SMOTE and Tuning

Model	Accuracy	Precision	Recall	F1-Score
RF	0.82	0.84	0.73	0.77
RF+S	0.81	0.77	0.76	0.77
RF+S+T	0.81	0.76	0.75	0.75

Table 5 shows that the Random Forest model performed differently at each experimental stage. Under the initial conditions (before applying SMOTE), the model achieved an accuracy of 0.82, a precision of 0.84, a recall of 0.73, and an F1-score of 0.77. The high precision value indicates the model's ability to minimize false positives, but the relatively low recall suggests limitations in detecting minority classes.

After applying SMOTE, the recall value increased to 0.76, while the F1-score remained at 0.77. This indicates that the model became more sensitive to minority classes. However, the precision decreased to 0.77, and the accuracy decreased slightly to 0.81, reflecting a trade-off due to the increased number of false positives.

Furthermore, after hyperparameter tuning, the model demonstrated more balanced performance, with a precision of 0.76, a recall of 0.75, and an F1-score of 0.75, while accuracy remained at 0.81. Although there was no improvement in accuracy or F1-score, the tuning process yielded a more stable model with better generalization to the test data.

Based on the confusion matrix in Figure 7 after applying SMOTE and hyperparameter tuning, the Random Forest model demonstrated quite good classification performance across all classes. In the Good class, the model correctly classified 63 data points, but still misclassified 15 to the Moderate class and 3 to the Unhealthy class. In the Moderate class, the model performed best, with 95 correct predictions and misclassifying 13 data points to the Good class and 3 to the Unhealthy class. Meanwhile, in the Unhealthy class, the model correctly classified 10 data points, misclassifying 5 to the Moderate class and 1 to the Good class.

These results indicate that the model has high accuracy in the majority class, especially Moderate, and an improved ability to detect minority classes, such as Unhealthy, after applying SMOTE. However, there are still significant misclassifications between the Good and Moderate classes, indicating similar feature patterns in the two classes. Overall, the combination of SMOTE and hyperparameter tuning is effective at improving the model's performance balance, especially by increasing sensitivity to minority classes without significant performance degradation.

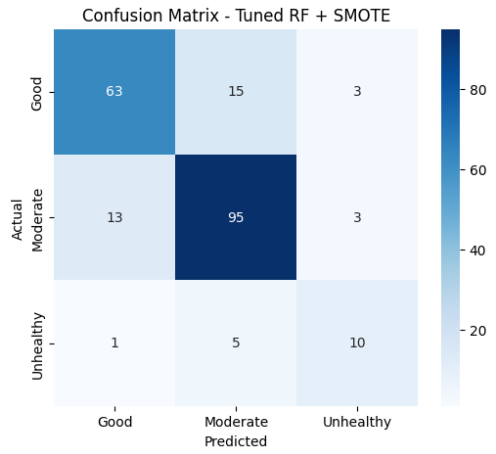


Figure 7. RF Confusion Matrix After SMOTE and Tuning

D. SHAP Analysis

To improve model transparency, this study applied SHapley Additive Explanations (SHAP) analysis using SHAP interaction value and mean SHAP interaction value analysis

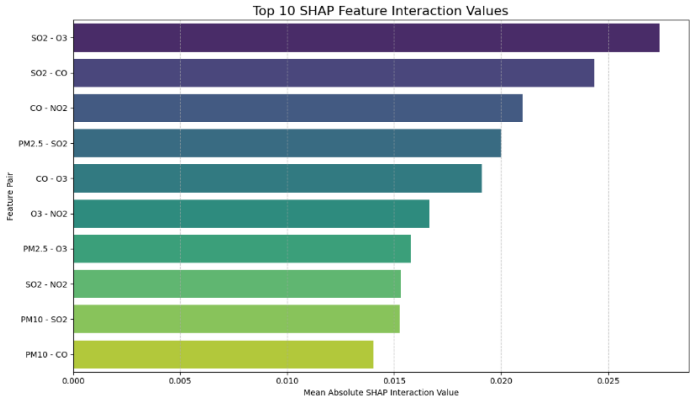


Figure 8. Mean SHAP Interaction Value

In this graph, the X-axis represents the value of the primary feature (pollutant concentration), which indicates the variation in the actual value of that feature in the dataset. The Y-axis shows the SHAP interaction value, which is the contribution of the interaction between two features to the model output, thus reflecting how much the combination of the two features influences the prediction results. The color gradient on the data points represents the value of the second feature interacting with the primary feature on the X-axis.

Based on these results, the interaction between pollutant parameters, such as SO_2 and O_3, emerged as one of the pairs with the highest contribution value, indicating that the combination of these two variables plays a significant role in determining air quality categories. The high mean SHAP interaction value for certain feature pairs indicates that the two features are not independent, but rather mutually reinforcing in influencing the model output.

Furthermore, the dominance of interactions between key pollutant features, such as CO, NO_2, and other pollutant gases, indicates that the model relies not only on the contributions of individual features but also on complex relationships between variables. This aligns with the multivariate nature of air quality data and the correlation between pollutants. Therefore, an explainable artificial intelligence-based approach using SHAP interaction can uncover patterns of relationships between features that were previously difficult to interpret in ensemble-based models like Random Forest.

Overall, this visualization provides a more comprehensive understanding of the model's decision mechanisms, where interactions between features are a key factor in improving accuracy and interpretability. These findings also reinforce the effectiveness of using mean SHAP interaction values for identifying the most influential feature pairs globally in air quality classification systems.

Further analysis using the mean SHAP value is shown in Figure 9.

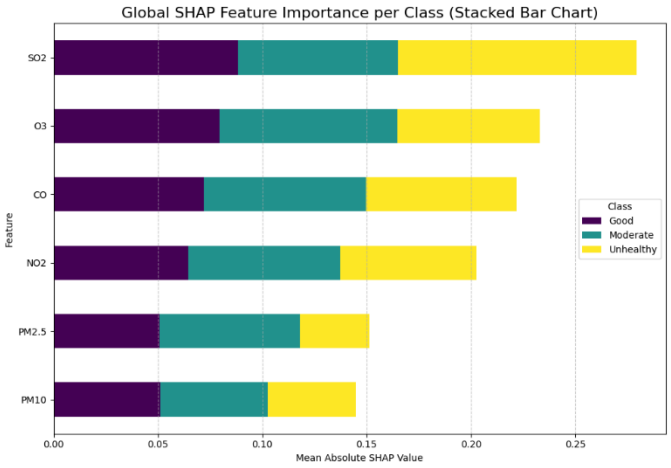


Figure 9. Mean SHAP Value

In Figure 9, the three-color diagram shows purple representing Class 0 (Good), green representing Class 1 (Moderate), and yellow representing Class 2 (Unhealthy).

The SO_2 parameter emerged as the feature with the most significant influence on model predictions across all classes. The length of the color bars in the graph indicates that the variance of SO_2 values carries the greatest weight in determining whether air quality is categorized as 'Good,' 'Moderate,' or 'Unhealthy.' Physically, this indicates that fluctuations in sulfur dioxide in the South Tangerang area are a key indicator of changes in urban air quality status.

O_3 ranks second, reflecting the role of secondary pollutants in urban air quality dynamics. The significant contribution of O_3 indicates that photochemical reactions in the South Tangerang atmosphere are quite intense, where pollutant precursors react with sunlight. O_3 is an important marker in the classification because it often spikes during the day when anthropogenic activity is at its peak.

Carbon monoxide ranks third. Although CO_2 is the dominant emission from private vehicles (the light transportation sector), its influence on model output is not as strong as SO_2 or O_3. This suggests that although CO_2 is present evenly in the atmosphere, fluctuations in its values in the 2020–2022 dataset tend to be stable and rarely act as a single trigger that drives air quality status to the 'Unhealthy' level.

In terms of emission sources, SO_2 is generally produced from the combustion of fossil fuels in the transportation and industrial sectors, which are dominant activities in urban areas and buffer zones such as South Tangerang. Therefore, the dominance of SO_2 in the model analysis results not only reflects statistical patterns but also aligns with actual environmental conditions in the study area [20].

E. Implementation and Deployment

fter the interpretability phase using SHAP was completed and the model was declared valid in terms of performance and transparent in terms of feature contribution, the next step was

implementing a cloud-based system using Hugging Face Spaces.

The deployment architecture consists of three main components. The trained RF classification model is saved in.pkl format. The application layer uses Python with the Gradio library as the web interface. The hosting layer uses the Hugging Face Spaces platform. The system interface is shown in Figure 11.

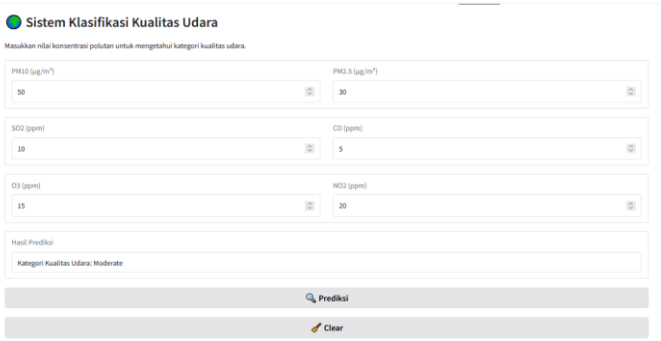


Figure 11. Air Classification System Interface

The Hugging Face deployment repository folder structure is as follows:

air-quality-classification
- app.py
- model_final_rf.pkl
- requirements.txt
- README.md

F. Analysis and Discussion

Based on the evaluation results, applying SMOTE to the Random Forest model did not yield a significant increase in accuracy. The accuracy only decreased marginally from 0.82 to 0.81 after oversampling and remained stable after hyperparameter tuning. This finding aligns with several studies that suggest that oversampling methods like SMOTE do not always improve accuracy, especially in ensemble models that are inherently robust to data imbalance.

However, the recall metric improved from 0.73 to 0.76 after applying SMOTE. This indicates that the model is more sensitive in detecting minority classes, which is the primary goal of the oversampling technique.

However, this increase in recall is accompanied by a decrease in precision from 0.84 to 0.77, indicating an increase in false positives. This phenomenon is a common trade-off in handling imbalanced data, as reported in various studies related to SMOTE-based classification [21].

The effectiveness of SMOTE in this study was further evaluated using Receiver Operating Characteristic (ROC) curves to measure the model's discriminatory ability for each category. Figure 12 shows that the post-optimization Random Forest model achieved an Area Under Curve (AUC) value of 0.94 for the 'Good' class, 0.93 for the 'Moderate' class, and 0.98 for the 'Unhealthy' class. The consistent AUC values above 0.90 across all classes also provide scientific justification that the SMOTE integration successfully addressed bias issues in the South Tangerang ISPU data, ensuring the model's high reliability for use in critical air quality classification systems..

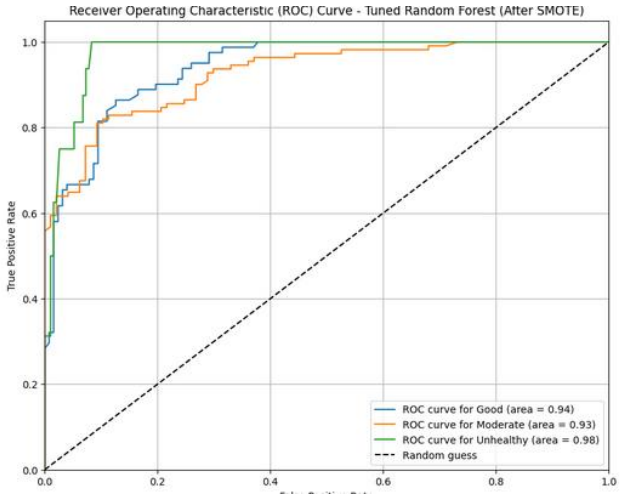


Figure 12. ROC Curve RF After SMOTE

Interpretability analysis using the Mean Absolute SHAP Value provides transparency into the decision-making mechanism of the Random Forest model, which is technically a black box. The main findings indicate that SO_2 is the most determinant feature in air quality classification in South Tangerang, followed by O_3, CO, and NO_2. The dominance of SO_2 in this SHAP analysis is spatially correlated with high fossil fuel combustion activity in the industrial and transportation sectors in urban buffer areas such as South Tangerang. This phenomenon aligns with the characteristics of the Greater Jakarta (Jabodetabek) region, which has high mobility and high fossil fuel use intensity. A study by Aulia et al. (2022) showed that increased transportation and industrial activity in the Greater Jakarta area significantly contribute to air pollutant emissions, including SO_2, resulting from fossil fuel combustion [22].

Furthermore, the SHAP visualization in stacked bar format shows that SO_2 is not only dominant globally but also a critical predictor of the 'Unhealthy' category (Class 2). This was validated by SHAP Interaction Value analysis, which demonstrated a synergistic non-linear relationship between SO_2 and other pollutants. When SO_2 concentrations increased alongside other parameters, the model's probability of assigning an 'Unhealthy' status increased significantly. This phenomenon demonstrates that the model does not rely solely on simple linear correlations but successfully captures the complex chemical dynamics of the urban atmosphere. The model's success in identifying SO_2 as a critical component aligns with theories of air pollution in industrial-urban areas, providing a

scientific basis for policymakers to prioritize sulfur emission control to reduce the frequency of days with poor air quality.

The final stage of this research was to transform the experimental model into a ready-to-use cloud-based classification system. Implementation on the Hugging Face Spaces platform using the Gradio framework demonstrated that the optimized Random Forest model is sufficiently computationally efficient for production environments. The system's three-layer architecture Model Layer (.pkl), Application Layer (Python & Gradio), and Hosting Layer (Cloud) ensures functional separation, facilitating maintenance and regular model updates without disrupting the user interface. The use of a modular repository structure, as seen in app.py and requirements.txt, ensures system scalability for future feature additions or data integration from wireless sensor sources. Although the system currently relies on manual entry of historical data, this successful deployment provides proof of concept that integrating transparent machine learning techniques (via SHAP analysis) with cloud computing can yield practical solutions for air pollution mitigation in urban areas. In the future, this system has the potential to be further developed through API (Application Programming Interface) integration with IoT sensor networks to enable real-time air quality monitoring.

IV. Conclusion

From the above results, the following conclusions can be drawn:

1. The Random Forest algorithm proved to be the superior model compared to SVM and KNN, achieving an accuracy of 0.81 and an F1-Score of 0.75. The use of SMOTE proved crucial in increasing the model's sensitivity to the minority (Unhealthy) class without sacrificing predictive stability in the majority class.
2. Analysis using SHAP (Explainable AI) revealed that SO_2 was the most important determining feature influencing air quality status at the study site, followed by O_3, CO, and NO_2. These findings provide a scientific basis for policymakers to prioritize sulfur emission control in industrial areas and heavy transportation.
3. The optimized model was successfully deployed to the Hugging Face Spaces cloud platform with a Gradio-based interface. This system provides fast, interactive classification services while demonstrating that complex machine learning models can be transformed into practical, publicly accessible applications.

Further development can focus on direct integration with IoT devices and on adding additional pollutant parameters to improve the system's accuracy and coverage. Exploration of using deep learning-based algorithms such as Long Short-Term Memory (LSTM), to predict future air quality is also underway.

V. Reference

[1] Farhatun Haya, Khaira Nisa, Rio Febrian Ladipasa, Ari Suriani, And Afriza Media, "Dampak Polusi Udara Terhadap Kesehatan Manusia," *WISSEN : Jurnal Ilmu Sosial Dan Humaniora*, Vol. 3, No. 2, 2025, Doi: 10.62383/Wissen.V3i2.753.

[2] I. E. Agbehadji and I. C. Obagbuwa, "Explainable Artificial Intelligence and Machine Learning for Air Pollution Risk Assessment and Respiratory Health Outcomes: A Systematic Review," 2025. Doi: 10.3390/Atmos16101154.

[3] N. Novitasari and M. K. Anwar, "Kinerja Dinas Lingkungan Hidup Dalam Pengendalian Pencemaran Udara di Kota Tangerang Selatan," *Restorica: Jurnal Ilmiah Ilmu Administrasi Negara Dan Ilmu Komunikasi*, Vol. 11, No. 2, Oct. 2025.

[4] Q. Liu, B. Cui, And Z. Liu, "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling," *Atmosphere (Basel).*, Vol. 15, No. 5, 2024, Doi: 10.3390/Atmos15050553.

[5] Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review," 2018. Doi: 10.3390/App8122570.

[6] A. A. Nababan, M. Jannah, M. Aulina, And D. Andrian, "Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu)," *Jtik (Jurnal Teknik Informatika Kaputama)*, Vol. 7, No. 1, 2023, Doi: 10.59697/Jtik.V7i1.66.

[7] I Gusti Ayu Nandia Lestari and I Komang Agus Ady Aryanto, "Peningkatan Akurasi Klasifikasi Kualitas Udara Melalui Oversampling Dengan Metode Support Vector Machine Dan Random Forest," *Jurnal Sistem Dan Informatika (JSI)*, Vol. 18, No. 1, 2023, Doi: 10.30864/Jsi.V18i1.596.

[8] T. S. Vausia and A. Kristiyanto, "Comparison Of Random Forest And Support Vector Machine Learning Algorithms In Sentiment Analysis Of Gojek User Reviews," *Jurnal Komtekinfo*, Vol. 4, No. 12, Pp. 239–245, 2025, Doi: Https://Doi.Org/10.35134/Komtekinfo.V12i4.669.

[9] R. Goran *Et Al.*, "Identifying And Understanding Student Dropouts Using Metaheuristic Optimized Classifiers And Explainable Artificial Intelligence Techniques," *IEEE Access*, Vol. 12, 2024, Doi: 10.1109/ACCESS.2024.3446653.

[10] A. Houdou *Et Al.*, "Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review," 2024. Doi: 10.4209/Aaqr.230151.

[11] A. S. Iffadah, Trimono, And Dwi Arman Prasetya, "Shapley Additive Explanations Interpretation of The Xgboost Model In Predicting Air Quality In Jakarta,"

Jurnal Riset Informatika, Vol. 7, No. 3, 2025, Doi: 10.34288/Jri.V7i3.366.

[12] G. L, N. S. Sriya, M. Sowmya, A. R. H. Lakshmi, And R. Amirtharajan, "Explainable AI For Urban Air Quality: SHAP Interpretation of Stacked Ensemble AQI Forecast," *Theor. Appl. Climatol.*, Vol. 156, No. 10, 2025, Doi: 10.1007/S00704-025-05741-3.

[13] Y. Choi, B. Kang, And D. Kim, "Utilizing Machine Learning-Based Classification Models for Tracking Air Pollution Sources: A Case Study in Korea," *Aerosol Air Qual. Res.*, Vol. 24, No. 7, 2024, Doi: 10.4209/Aaqr.230222.

[14] N. El Furqany, "Optimizing Air Quality Index Classification Using Multiple Machine Learning Models and Oversampling Techniques," *International Journal of Artificial Intelligence In Medical Issues*, Vol. 3, 2025.

[15] G. C. V. Putri, E. Saputera, And D. P. Kesuma, "Systematic Literature Review: Pemanfaatan Cloud Computing Dalam Pengembangan Kecerdasan Buatan," *Buletin Sistem Informasi Dan Teknologi Islam*, Vol. 6, No. 3, 2025, Doi: 10.33096/Busiti.V6i3.2813.

[16] Dinas Lingkungan Hidup Tangerang Selatan, "Indeks Kualitas Air Dan Indeks Kualitas Udara di Kota Tangerang Selatan Tahun 2022," Https://Data.Tangerangselatankota.Go.Id.

[17] E. P. K. Sari, Y. D. Rosita, And S. Khomsah, "Implementasi Model Random Forest Untuk Klasifikasi Kualitas Udara Di Jakarta Periode 2019-2024," *Journal Of Information Technology Literacy*, Vol. 1, No. 1, Pp. 1–6, 2025.

[18] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, And M. Ismail, "SMOTE For Handling Imbalanced Data Problem : A Review," In *2021 6th International Conference On Informatics And Computing, ICIC 2021*, 2021. Doi: 10.1109/ICIC54025.2021.9632912.

[19] Prasetyo, D., Mahardika, F., Hariyadi, H., Nurhayati, N., Kurniabudi, K., Permadi, A., ... & Jabbar, M. S. A., *Cloud Computing*. Yogyakarta: Penamuda, 2024.

[20] D. D. Lestiani *Et Al.*, "Selected Elements Characterization of Fine Particulate Matter PM2.5using Synchrotron Radiation XRF," In *AIP Conference Proceedings*, 2021. Doi: 10.1063/5.0066285.

[21] M. Buda, A. Maki, And M. A. Mazurowski, "A Systematic Study of The Class Imbalance Problem in Convolutional Neural Networks," *Neural Networks*, Vol. 106, 2018, Doi: 10.1016/J.Neunet.2018.07.011.

[22] Emir Aulia, Y. Chadirin, And A. Pribadi, "Analisis Sebaran SO2 Pada Musim Wabah Covid-19 Menggunakan Satelit Aura di Wilayah Jabodetabek," *Jurnal Teknik Sipil Dan Lingkungan*, Vol. 7, No. 2, 2022, Doi: 10.29244/Jsil.7.2.113-128.