PREDICTION OF AIR POLLUTION STANDARD INDEX (ISPU) CATEGORIES IN DKI JAKARTA USING THE GRADIENT BOOSTING ALGORITHM

Aditya Pratama[1], Nurman Hariyanto[2], Adha Maulana[3], Denny Primanda[4], Awanis Hidayati[5] , Wahyu Prayitno[6]
Nahdlatul Ulama University Of West Kalimantan[1,2,3,4,5,6]
Email: adityapratamabadra@unukalbar.ac.id

Abstract
Air quality in DKI Jakarta is a significant problem that threatens the health of its people. To give the public an idea of the level of air pollution in certain areas and the risks it poses, there is the Air Pollution Standard Index (ISPU) which provides the relevant information. The objective of this study is the development of the ISPU category prediction model in DKI Jakarta based on the Gradient Boosting algorithm. The ISPU data used includes the concentrations of PM10, PM2.5, SO2, CO, and O3. The developed Gradient Boosting model was trained on the historical ISPU data and then its performance was assessed on the basis of accuracy, precision, recall and F1-score metrics. In this study, the Gradient Boosting model has been demonstrated to be able to predict ISPU categories with an improved degree of accuracy. It has a potential purpose – the provision of early warning with respect to ISPU categories to allow the general public to take measures that would reduce the degree of air pollution they are exposed to. This research adds up to the establishment of an air quality early warning system, which is entirely instrumental in the advancement of the health standards of the people living in the DKI Jakarta region.

Keywords: *ISPU, Gradient Boosting, prediction, air quality, DKI Jakarta*

A. Introduction

Air quality in urban areas, especially in DKI Jakarta, is a significant concern due to its impact on public health and the environment. Air pollution can cause various health problems, such as respiratory diseases, cardiovascular diseases, and cancer (Syuhada et al., 2023). To measure the level of air pollution and provide information to the public, the Air Pollution Standard Index (ISPU) is used. ISPU is a number that indicates the level of air quality and its impact on health Istiana et al., 2023)..

ISPU category prediction is vital to provide early warning to the public so that they can take preventive measures to reduce the impact of air pollution. This study aims to build a prediction model for the ISPU category in DKI Jakarta using the Gradient Boosting algorithm. *Gradient Boosting* is a robust machine learning algorithm that has been proven effective in various prediction tasks. This algorithm works by combining several simple prediction models (decision trees) to produce a more accurate model (Toharudin et al., 2023).

The data used in this study is historical data from the DKI Jakarta ISPU, which includes the concentrations of PM10, PM2.5, SO2, CO, and O3. These features were chosen because they are the main parameters in calculating the ISPU. [6] The Gradient Boosting model will be trained using historical data, and its performance will be evaluated using accuracy, precision, recall, and F1-score metrics. This research is expected to contribute to the development of an early warning system for air quality in DKI Jakarta. The resulting prediction model can be used to provide information to the public about the ISPU category in the future so that necessary preventive measures can be taken. In addition, this research can also provide information to the government to take appropriate policies in controlling air pollution.

B. Literature Review and Hypothesis Development

Air pollution has become a critical environmental issue globally, impacting human health and the economy. Studies leveraging machine learning for air quality prediction have focused not only on enhancing predictive accuracy but also on improving the interpretability and real-world applicability of models (Houdou et al., 2024). Various approaches have been developed, each with unique strengths and contributions to understanding air quality and informing policy.

1. Machine Learning in Air Quality Prediction

Machine learning algorithms have become popular tools for air quality prediction due to their ability to process large datasets and identify complex patterns. For instance, Sridevi (2023) developed a machine learning model using Gradient Descent-boosted multivariable regression to forecast India's Air Quality Index (AQI), achieving 96% accuracy. The model's efficiency was further enhanced through the application of cost estimation, and additional support from XGBoost and Light GBM algorithms improved its robustness and predictive power. This demonstrates how advanced algorithms can outperform traditional models by integrating parameter-reducing formulations (Sridevi, 2023).

In another study, a systematic review by Houdou et al. (2024) highlighted the use of interpretable machine learning models in air pollution prediction. Among 5,396 identified studies, 480 focused on air pollution prediction, and 56 provided model interpretations. Methods such as Shapley Additive Explanations (SHAP) and Partial Dependence Plots (PDP) were frequently used to identify influential features, enhancing understanding and making machine learning outcomes accessible to non-experts. SHAP was noted as the most popular method due to its comprehensive feature importance analysis (Houdou et al., 2024).

2. Impact of Air Pollution on Health and Economy

Syuhada et al. (2023) provided an in-depth analysis of the health and economic burdens of air pollution in Jakarta. Their study quantified the impacts of fine particulate matter (PM2.5) and ground-level ozone (O3), revealing over 7,000 adverse health outcomes in children and more than 10,000 deaths annually linked to air pollution. The total annual cost of these health impacts was estimated at approximately USD 2.94 billion. This economic assessment emphasized the need for proactive measures to improve air quality and protect public health, using localized data to better tailor policies (Syuhada et al., 2023).

3. Applications of Advanced Machine Learning Algorithms

Gradient Boosting algorithms have also been explored for air quality prediction due to their flexibility and strong predictive performance. A study conducted on air pollution in Jakarta employed Gradient Boosting to predict the Air Pollution Standard Index (ISPU) categories, showing promising results with high accuracy and comprehensive evaluation metrics such as precision, recall, and F1-score (User's Study). The study underscored the potential of Gradient Boosting in providing early warnings, which could help the public take preventative measures against exposure to air pollution.

4. Importance of Interpretability in Machine Learning

The balance between model accuracy and interpretability is crucial, as highlighted by Houdou et al. (2024). Model-agnostic methods like SHAP are valuable tools that provide insights into which pollutants most affect air quality predictions. This interpretability can bridge the gap between complex model outcomes and actionable insights for policymakers and health officials. Similarly, Sridevi's (2023) work, although focused on accuracy, showcases how parameter optimization and multi-algorithm integration contribute to both performance and model comprehension.

The body of research underscores the importance of using machine learning for air quality prediction, integrating high accuracy with interpretability to aid decision-makers. While studies like those by Sridevi (2023) and Syuhada et al. (2023) focus on developing predictive models and quantifying economic impacts, Houdou et al. (2024) and user-specific studies stress the value of transparent models that enhance public understanding and policy formation. Future research should continue to develop models that balance accuracy with interpretability and use findings to drive effective environmental and public health interventions.

C. Research Method

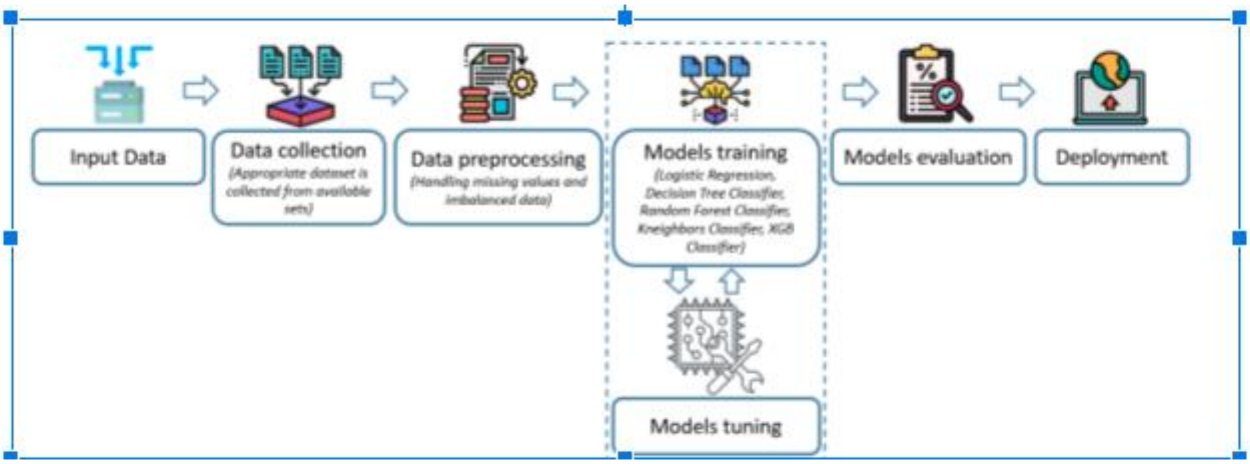


Figure 1. Workflow for Air Quality Classification

1. Input Data

The first stage is to determine relevant input data for this research. The data used includes air pollution parameters such as PM10, PM2.5, SO2, CO, O3, and NO2 obtained from air quality monitoring stations in DKI Jakarta as well as other environmental data sources.

2. Data Collection

The data collected must cover a specific period (for example, 2010–2023) so that the model has enough historical data to make accurate predictions. This process involves:

- Access data from trusted sources such as the Environment Agency and air quality monitoring stations.
- Ensure that the data includes all parameters required for the calculation of the Air Pollution Standard Index (ISPU).

3. Data Preprocessing

The data that has been collected needs to be processed to improve quality and ensure the accuracy of model predictions. The steps in preprocessing include:

- **Handling Missing Values**: Using imputation techniques such as K-Nearest Neighbors (KNN) or interpolation to fill in missing data.

- **Handling Imbalanced Data**: Applying resampling techniques such as oversampling with SMOTE (Synthetic Minority Over-sampling Technique) if there is a class imbalance in the data.
- **Normalization/Standardization**: Normalize or standardize data so that features are on a uniform scale.
- **Feature Engineering**: Adds additional relevant features such as humidity, temperature and wind speed if available.

4. Models Training

The model training process involves several machine learning algorithms to determine the best model. The algorithm used Gradient Boost**.**The model will be trained using training data (80% of the dataset) and optimized to achieve the best accuracy and performance.

5. Models Tuning

To improve model performance, parameter tuning is carried out using techniques such as:

- Grid Search or Random Search to find the optimal parameter combination.
- Use of cross-validation (e.g., 5-fold) to avoid overfitting and ensure a stable model.

6. Models Evaluation

The trained model is evaluated using the following metrics:

- **Accuracy**: Shows how well the model classifies ISPU categories.
- **Precision, Recall, F1-Score**: Assess the model's performance in correctly classifying categories, especially minority classes.
- **ROC-AUC Score**:Evaluate the model's ability to differentiate between positive and negative categories overall.

7. Deployment

After the model is selected, the next step is to deploy the model into a real-time monitoring system. This model will be integrated with a platform that can receive real-time data from air quality sensors in Jakarta and provide ISPU predictions automatically. This system can be used by related parties to provide early warnings and support more effective environmental policies.

This research method covers all stages from data collection to deployment, with a focus on comprehensive data processing and selecting the best model optimized for accuracy and interpretability. The deployment of this model can make a significant contribution to the air quality early warning system in DKI Jakarta and help improve public health.

D. Discussion

The dataset used in this research, consisting of 4,626 daily observations related to air quality in Jakarta over several years, has undergone extensive cleaning and processing to ensure its usability for machine learning models. Due to the presence of missing values and inconsistencies in the raw data, it was processed into a simpler and more comprehensible format. While some columns are straightforward, others required specific assumptions and modifications to align with the research objectives.

Each row in the dataset represents an observation for a single day, while the columns reflect different air pollution parameters influencing air quality. Most missing values were found in pollutant concentration columns, such as PM10, PM2.5, and other pollutants. To handle these missing values, the median value of each column was used as a replacement to maintain data integrity and provide a complete, consistent dataset for machine learning models.

The target variable in the dataset is the air quality category, which the model aims to predict. These categories include "Baik" (Good), "Sedang" (Moderate), "Tidak Sehat" (Unhealthy), and "Sangat Tidak Sehat" (Very Unhealthy). Categories with minimal representation, such as "Dangerous" and "No Data," were removed during preprocessing due to insufficient data for effective model training.

Numerical columns representing concentrations of pollutants such as PM10, PM2.5, SO2, CO, O3, and NO2 were standardized to ensure uniform scaling for the model. Categorical columns were processed using label encoding, converting them into numerical values—for instance, "Baik" was mapped to 0, "Sedang" to 1, and so on.

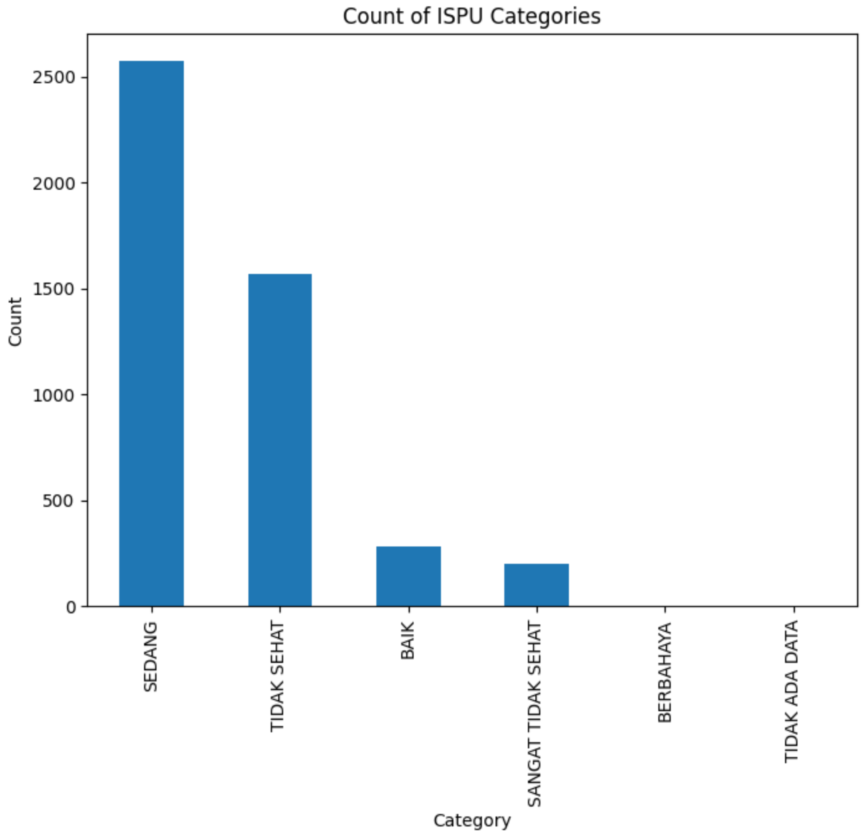
Initial analysis of the distribution of target variables showed that most data fell under the "Tidak Sehat" category, while categories such as "Baik" and "Sedang" had comparatively fewer data points. This class imbalance can affect model performance, suggesting the need for oversampling techniques or class weighting to enhance model efficacy. Overall, this dataset has been thoroughly processed, ensuring its quality for use in machine learning tasks and providing valuable insights into long-term trends and seasonal patterns in air quality. A detailed explanation of each attribute in the dataset is shown in Table 1.

Attribute Name	Description	Data Type	Function
tanggal	Air quality observation date, formatted in the standard form YYYY-MM-DD.	Date/Time	Temporal analysis, such as identifying seasonal or daily trends in air quality data.
pm10	Concentration of air particles with a diameter of ≤10 micrometers, measured in µg/m³.	Numeric	The main parameters for assessing air quality; High concentrations can cause respiratory problems.
pm25	Concentration of air particles with a diameter of ≤2.5 micrometers, measured in µg/m³	Numeric	The main indicator of air pollution that can penetrate deep into the human respiratory system.
so2 (Sulfur Dioksida)	Konsentrasi gas sulfur dioksida di udara, diukur dalam satuan µg/m³.	Numeric	The concentration of sulfur dioxide gas in the air, measured in µg/m³
co (Karbon Monoksida)	The concentration of carbon monoxide gas in the air, measured in mg/m³.	Numeric	Toxic gases from incomplete combustion; High levels reduce the blood's capacity to transport oxygen..
o3 (Ozon)	Concentration of ozone gas in the air, measured in µg/m³.	Numeric	Secondary pollutants formed from chemical reactions; affecting human health and vegetation
no2 (Nitrogen Dioksida)	The concentration of nitrogen dioxide gas in the air, measured in µg/m³.	Numerik	Pollutants from motor vehicle emissions; High levels can cause respiratory problems..
categori	Air quality categories are based on the concentration of main pollutant parameters.	Kategorikal	Grouping air quality into classes such as "Baik," "Sedang," "Tidak Sehat," and "Sangat Tidak Sehat."

With these preparations, the dataset is now prepared for further analysis using machine learning algorithms, enabling more precise results for air quality classification.

Data Visualization

In this section, various features within the dataset are visualized to identify patterns, correlations between parameters, and their connection to air quality categories (categories). These visualizations offer valuable insights into the data structure and support the development of machine learning-based predictive models.



1. Distribution of Air Quality Categories

As depicted in Figure 2, the distribution of air quality categories reveals that the majority of the data falls into the "Moderate" and "Unhealthy" categories, whereas the "Good" and "Very Unhealthy" categories have significantly fewer data points. This imbalance suggests that the dataset is predominantly composed of the majority class, potentially impacting the machine learning model's performance, particularly in predicting the minority class. To address this, techniques such as oversampling or applying special weighting to minority classes may be required to enhance the model's accuracy for underrepresented categories.

Figure 2. Distribution of Air Quality Categories

2. Relationship between PM10 and PM2.5

The relationship between PM10 and PM2.5 concentrations is illustrated in Figure 3. This scatter plot reveals a strong positive correlation between these two parameters, showing that as PM10 levels increase, PM2.5 levels also rise consistently. This suggests that these particles often share common sources, such as vehicle emissions or fossil fuel combustion. This correlation is noteworthy as a high degree of similarity between features can impact model performance, particularly in algorithms that are sensitive to multicollinearity.

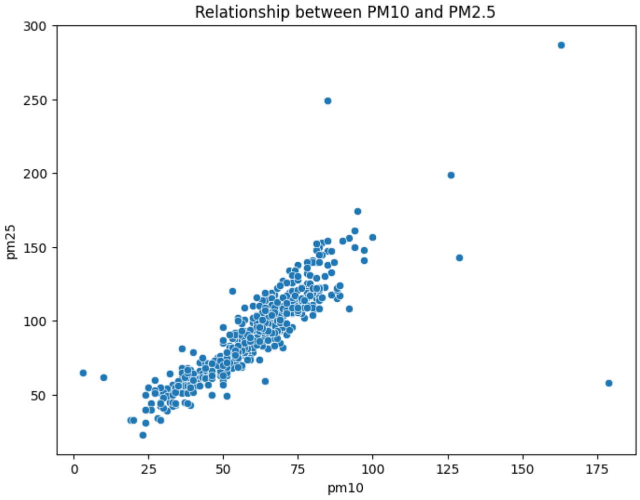


Figure 3 Relationship Between PM 10 and PM2.5

3. Trends in PM10 Concentrations over Time

Trends in PM10 concentrations over time are illustrated in Figure 4, showcasing the changes from 2010 to 2023. The graph highlights seasonal variations, with distinct peaks in concentration during certain periods. These patterns indicate that air pollution levels are affected by human activities and specific weather conditions. This insight is essential for considering the time factor as a variable in achine learning models, particularly for time-based predictive analyses.

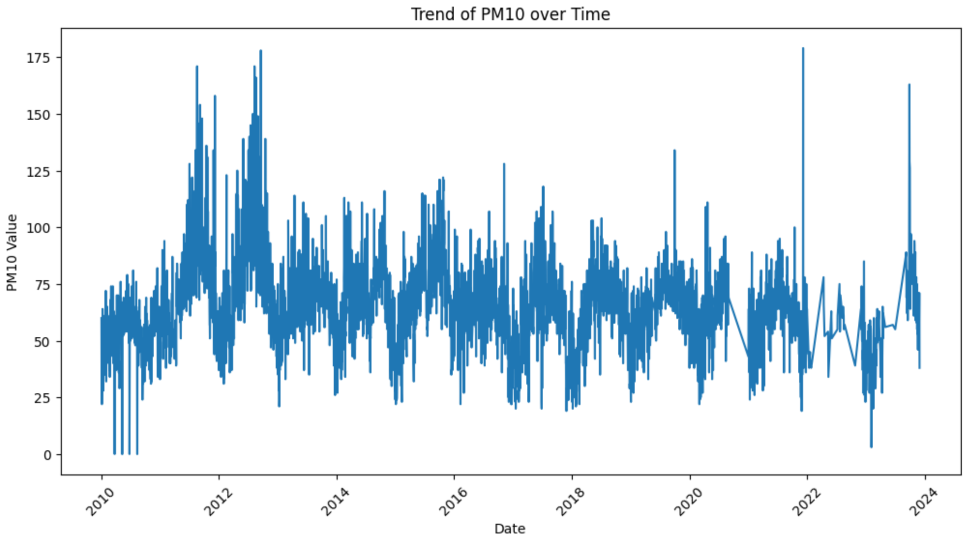


Figure 4. Trend of PM10 over time

4. Pollutant Concentration Distribution

The distribution histogram for each pollutant parameter offers a snapshot of concentration value distributions in the dataset. Key findings from each histogram are as follows:

1. **PM10 (Figure 5)**: The majority of PM10 concentrations fall within the 50–100 µg/m³ range, indicating the prevalent presence of coarse particles in the air.
2. **PM2.5 (Figure 6)**: Most PM2.5 concentrations lie between 75–125 µg/m³, suggesting that fine particle levels are higher than PM10, which aligns with typical urban air pollution characteristics.
3. **SO2 (Sulfur Dioxide) (Figure 7)**: The distribution of SO2 shows two distinct peaks, potentially pointing to two primary emission sources, such as industrial activities and vehicular traffic.
4. **CO (Carbon Monoxide) (Figure 8)**: CO concentrations mostly range from 20–40 mg/m³, displaying a distribution close to normal.
5. **O3 (Ozone) (Figure 9)**: The distribution of ozone is skewed, with lower concentrations being more frequent. This pattern aligns with ozone's nature as a secondary pollutant formed through atmospheric chemical reactions.

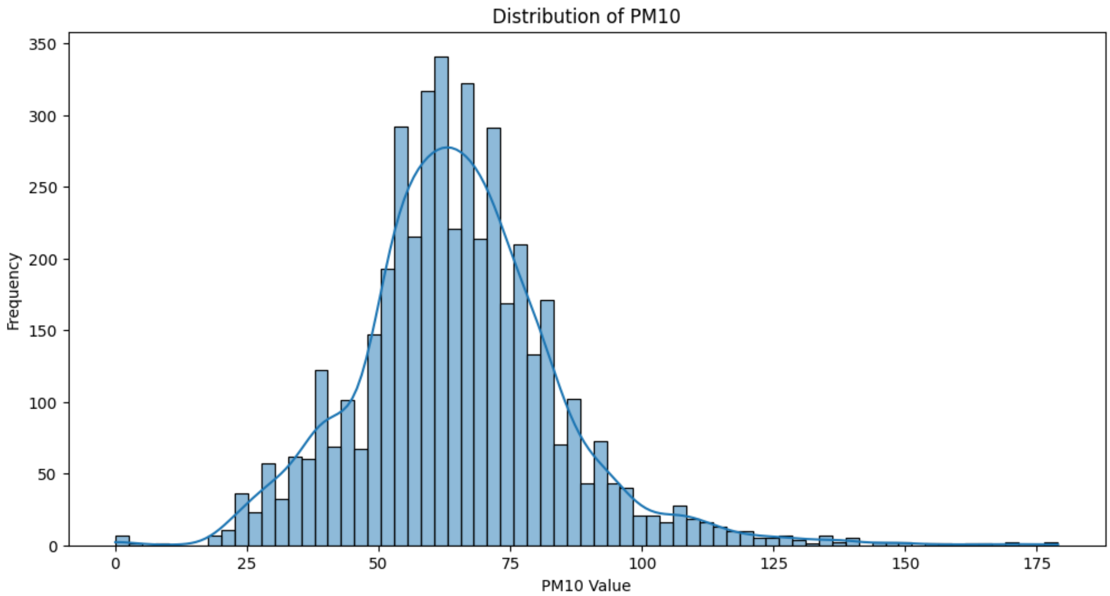


Figure 5. Distribution of PM10

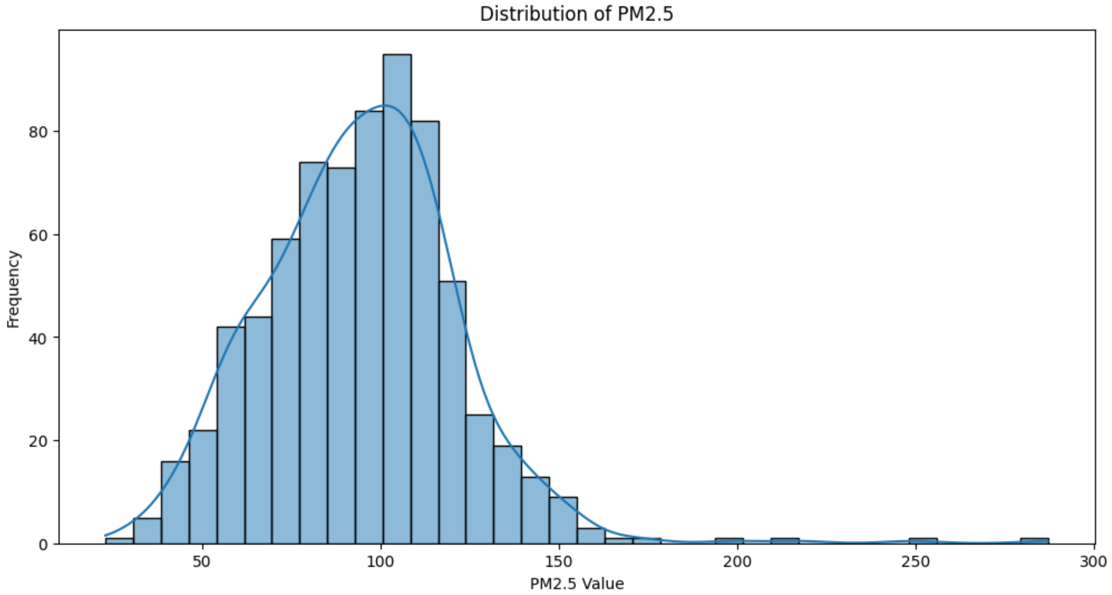


Figure 6. Distribution of PM2.5

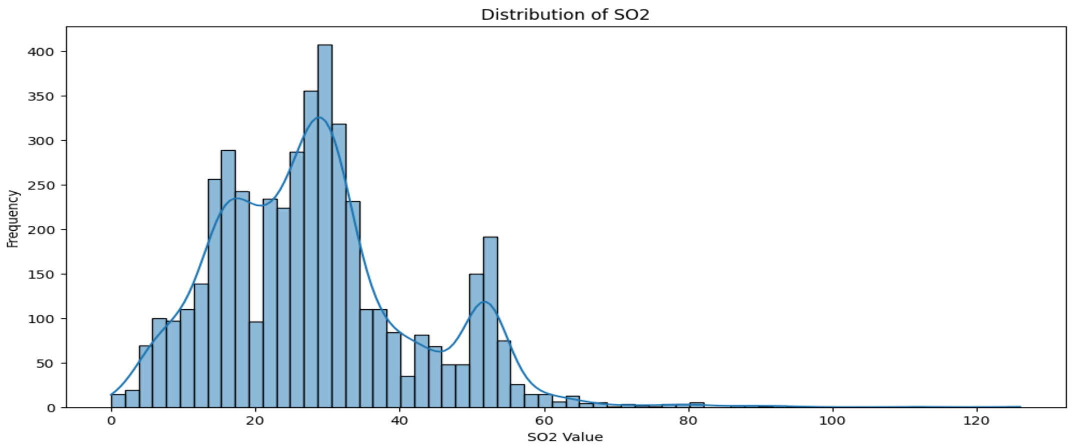


Figure 7. Distribution of SO2

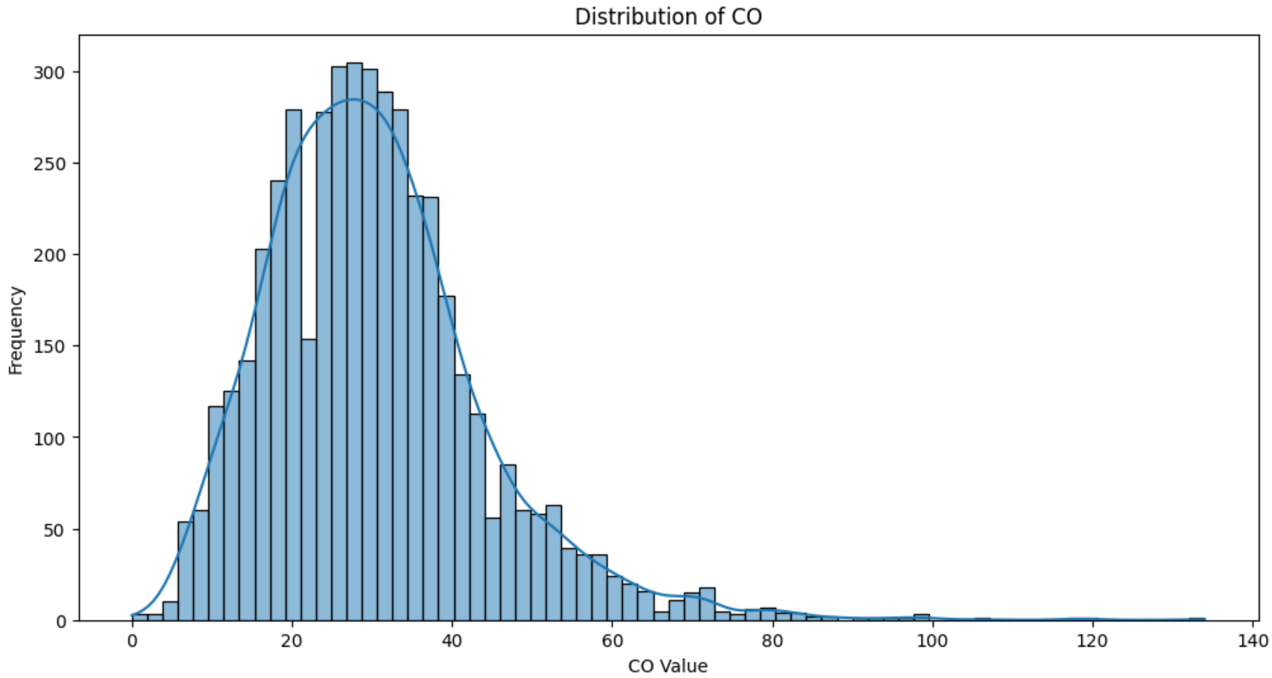


Figure 8. Distribution of CO

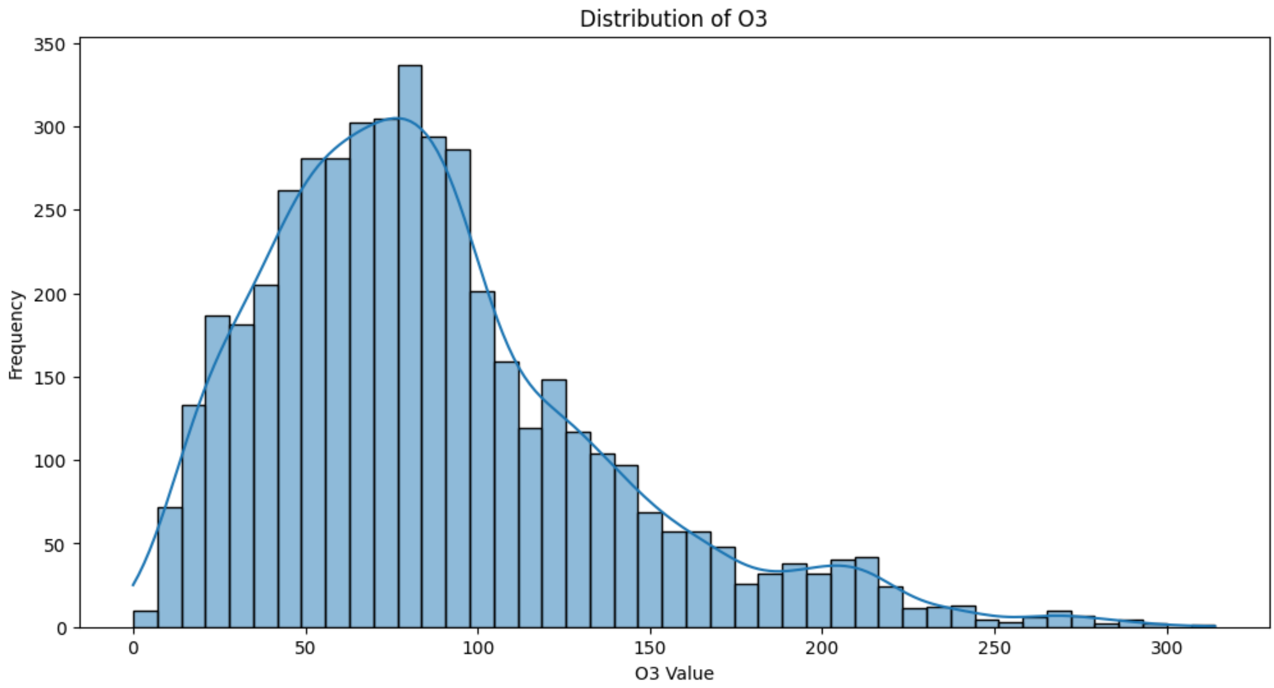


Figure 9. Distribution of O3

Based on the data visualization, several key conclusions can be drawn. First, there is an imbalance in the air quality categories within the dataset, with the majority class dominating, which necessitates special treatment to enhance model accuracy for minority classes. Second, the strong correlation between parameters such as PM10 and PM2.5 highlights their significance in air quality analysis. Third, the temporal patterns observed in PM10 concentrations underscore the importance of incorporating time as a factor in the analysis. Lastly, the distribution of pollutants like SO2, CO, and O3 offers valuable insights into the primary sources of urban air pollution. The next step involves leveraging these insights to select the most relevant features and optimize data preprocessing before implementing machine learning algorithms. If required, data normalization or handling of outliers can be conducted to further improve model prediction quality..

The application of the Gradient Boost algorithm to the air quality prediction dataset in DKI Jakarta shows very satisfactory results with an overall accuracy of **99.68%**, as seen in Figure 10.

```
Accuracy: 0.9967602591792657
                    precision    recall  f1-score   support

              BAIK       0.95      1.00      0.98        61
SANGAT TIDAK SEHAT       1.00      1.00      1.00        51
            SEDANG       1.00      0.99      1.00       531
       TIDAK SEHAT       1.00      1.00      1.00       283

          accuracy                           1.00       926
         macro avg       0.99      1.00      0.99       926
      weighted avg       1.00      1.00      1.00       926
```

Figure 10. Accuracy

The analysis of the results can be explained based on metric evaluation and confusion matrix visualization as follows:

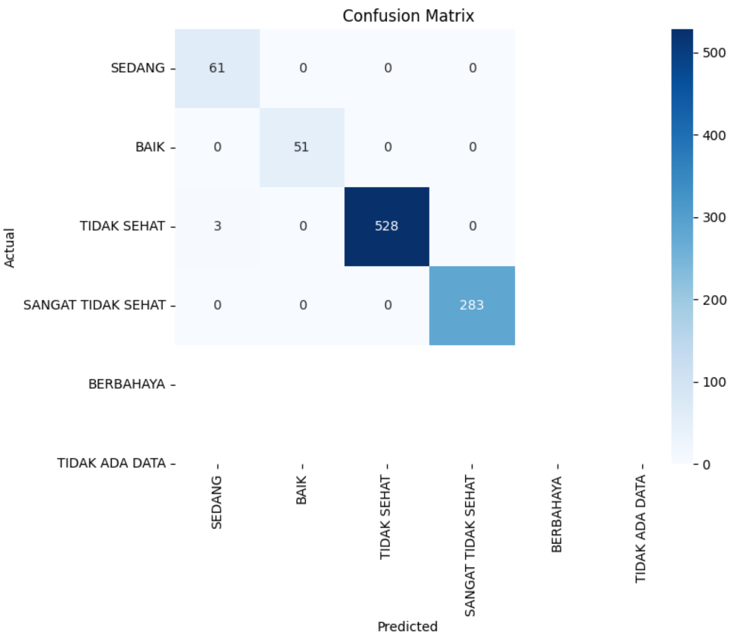


Figure 11. Confusion Matrix

1. Confusion Matrix Analysis:

- **'Baik' Category**: The model successfully classified all 61 actual data in the 'Baik' category with perfect precision, without any misclassification errors.

- **'Sedang' Category**: Of the 531 actual data, 528 were correctly classified as Sedang, and only 3 data were incorrectly classified into other categories.
- **'Tidak Sehat' Category**: All 283 actual data were successfully classified correctly by the model.
- **'Sangat Tidak Sehat' Category**: The model also showed perfect performance in classifying this data, with all 51 actual data points correctly predicted.

No data was classified into the 'Berbahaya' and 'Tidak Ada Data' categories because this data was removed in the preprocessing stage.

2. Precision, Recall, and F1-Score:

- **Precision**:The model shows high precision in all categories, with values ranging from 0.95 to 1.00, indicating that the model predictions are highly accurate for each category.
- **Recall**: The model has almost perfect recall, reflecting the model's ability to capture almost all of the actual data in each category.
- **F1-Score**:High F1-score values in all categories indicate a good balance between precision and recall, which means this model is very effective in predicting each category.

3. Macro dan Weighted Average:

E. **Macro Average**:The macro average precision, recall and F1-score values of 0.99 indicate consistent model performance in all categories, even though there is an imbalance in the amount of data between categories.
F. **Weighted Average**: The weighted average precision, recall, and F1-score are each 1.00, indicating that the model performs very well overall, especially for categories with a larger amount of data.

The results of applying the Gradient Boost algorithm show that this model has very good performance in classifying ISPU categories in DKI Jakarta. The model's near perfect performance in predicting categories such as Baik, 'Sedang', 'Tidak Sehat', and 'Sangat Tidak Sehat' strengthens the validity of the model for use in air quality prediction systems. However, special attention needs to be paid to minority categories that may require oversampling techniques or special weighting to ensure model robustness when applied to new data.

G. Conclusion

This research succeeded in developing a prediction model for the Air Pollution Standard Index (ISPU) category in DKI Jakarta using the Gradient Boosting algorithm, which shows high accuracy performance as well as almost perfect precision, recall and F1-score evaluation metrics. This model has proven effective in predicting air quality categories such as 'Baik', 'Sedang', 'Tidak Sehat' and 'Sangat Tidak Sehat', so it can be used as part of an early warning system to provide important information to the public regarding air quality conditions.

With this predictive model, people can take preventive action to reduce exposure to air pollution. This research contributes to improving the air quality monitoring system that supports environmental and public health policies in DKI Jakarta. Further implementations involving real-time data integration and adjustment for minority classes can be carried out to improve the robustness and accuracy of the model.

Bibliography

Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S., Siregar, A., Kusuma, R., Driejana, R., Ingole, V., Kass, D., & Mehta, S. (2023). Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, 20. https://doi.org/10.3390/ijerph20042916.

Istiana, T., Kurniawan, B., Soekirno, S., Nahas, A., Wihono, A., Nuryanto, D., Adi, S., & Hakim, M. (2023). Causality analysis of air quality and meteorological parameters for PM2.5 characteristics determination: Evidence from Jakarta. *Aerosol and Air Quality Research*. https://doi.org/10.4209/aaqr.230014

Toharudin, T., Caraka, R., Pratiwi, I., Kim, Y., Gio, P., Sakti, A., Noh, M., Nugraha, F., Pontoh, R., Putri, T., Azzahra, T., Cerelia, J., Darmawan, G., & Pardamean, B. (2023). Boosting algorithm to handle unbalanced classification of PM2.5 concentration levels by observing meteorological parameters in Jakarta-Indonesia using AdaBoost, XGBoost, CatBoost, and LightGBM. *IEEE Access, 11*, 35680–35696. https://doi.org/10.1109/ACCESS.2023.3265019

Houdou, A., Badisy, I., Khomsi, K., et al. (2024). *Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review*. Aerosol and Air Quality Research. https://doi.org/10.4209/aaqr.230151.

Sridevi, E. (2023). *Air Quality Monitoring Using Machine Learning Techniques*. International Journal of Scientific Research in Engineering and Management. https://doi.org/10.55041/ijsrem25321.

Syuhada, G., Akbar, A., Hardiawan, D., et al. (2023). *Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia*. International Journal of Environmental Research and Public Health, 20. https://doi.org/10.3390/ijerph20042916.