Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Prediction of the Air Quality Index in DKI Jakarta Province Using the CatBoost Method

Yoga Dwi Prasetyo[1], Fitria Nur Rahmadani[2], Mohammad Idhom[3]*, Trimono[4]

[1,2,3,4]*Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur*
22083010055@student.upnjatim.ac.id[1] , 22083010059@student.upnjatim.ac.id[2] , idhom@upnjatim.ac.id[3], trimono.stat@upnjatim.ac.id[4]*

Abstract

Air pollution in major cities like Jakarta continues to worsen due to various contributing factors, including unregulated industrial emissions, open waste burning, and the increasing number of private vehicles. This study aims to classify air quality levels based on the Air Pollution Standard Index (ISPU) using the CatBoost Classifier algorithm. The dataset comprises ISPU data from 2021 to 2024 sourced from Jakarta's public data portal, including parameters such as PM10, PM2.5, SO2, CO, O3, and NO2. After preprocessing and feature selection, the model was trained and evaluated using standard classification metrics. The CatBoost Classifier achieved high performance in major categories like "BAIK", "SEDANG", and "TIDAK SEHAT" with F1-scores exceeding 0.94. However, the "SANGAT TIDAK SEHAT" category could not be predicted accurately due to class imbalance. To address this, a hybrid model incorporating rule-based logic was employed, enabling accurate classification in the case of extreme pollution. The model also offers station-level predictions, supporting spatial analysis and early warning systems. The results demonstrate that the proposed approach provides a robust framework for air quality classification and real-time environmental monitoring.

Keywords: *Air Pollution, CatBoost, Classification, ISPU, Rule-based System*

1. Introduction

The issue of air quality has always been a special concern for local governments with the increasingly rapid industrial development in big cities today. Day by day, air quality is deteriorating due to high levels of air pollution, which often occur in large cities in Indonesia, such as Jakarta. The increase in air pollution can be caused by several things, such as the management of air waste from the industrial sector that does not follow local government rules or regulations, the habit of residents who like to burn garbage, and an increase in the number of private vehicle drivers.

The air in the universe contains several compound pollutants such as particulates (PM 10 and PM 2.5), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and nitrogen dioxide (NO2). High levels of these elements in a compound can affect public health, particularly in cases of respiratory diseases, which can even be fatal [1]. With accurate classification modeling, it can help the government to make appropriate policies that do not harm anyone. The presence of appropriate policies aims to control pollution so that air quality can meet the standards set by the government in the Regulation of the Minister of Environment and Forestry Number 14 of 2020 concerning Ambient Air Quality Standards [2].

Data mining techniques function to obtain the required information quickly. One of the data mining methods used in predicting is classification. In classifying air quality, there have been many studies that apply data mining techniques, one of which is research on air quality prediction using the KNN (K-Nearest Neighbor) algorithm to determine the optimal parameters in the dataset. The results of the study using this method, the author tested the values of K = 3 to K = 7, and it was found that the value of K = 7 had the best performance with the highest accuracy of 96%, precision of 92%, recall of 95%, and f-measure of 93%[3].

Furthermore, research that classifies the level of air quality in DKI Jakarta uses the Naïve Bayes algorithm. The result of this research is to create a test model for classification using the Naïve Bayes algorithm, which aims to find good results. The results of the classification on the data used are with an average accuracy of 88%, precision 85%, recall 96%, and f1-score 99% [4].

From the two previous studies that have been described to identify factors that influence air quality. This study will use the CatBoost Classifier algorithm to classify or predict the Air Quality Index in DKI Jakarta Province. The CatBoost Classifier algorithm has been tested and proven to work in air quality categorization modeling. The DKI Jakarta Provincial Government provides various types of data, one of which is the Air Pollution Standard Index (ISPU) data from 2021 to 2024, which will be used in this study. However, the dataset shows an imbalance, especially in the “VERY UNHEALTHY” category. To handle this, a hybrid approach is applied using class weighting and a rule-based system. Class weighting gives more importance to minority classes, while the rule-based logic ensures that if PM10 > 350 or PM2.5 > 300, the data is automatically labeled as “VERY UNHEALTHY.” This method aims to improve classification accuracy, especially in rare but critical cases

2. Research Methods

Research methods are steps that can be used to solve a problem that occurs in conducting research by following the provisions made by researchers so that the desired goals are achieved and obtain test results on the problems raised [5]. The following is the flow of the research method used, as shown in Figure 1.

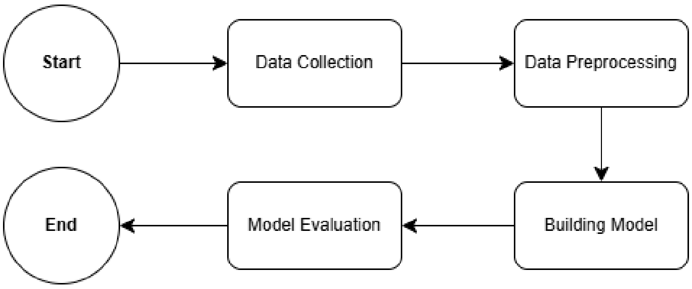

Figure 1: Research Flow Diagram

2.1. Data Collection

The first step in this study is data collection. The data used in this study is the Air Pollution Standard Index (ISPU) data for the period 2021 to 2024, which is accessed via the Satu Data Jakarta website https://satudata.jakarta.go.id/home. This data contains the Air Pollution Standard Index measured based on 5 air quality monitoring stations consisting of 12 attributes: data_period, date, station, PM10, PM2.5, SO2, CO, O3, NO2, max, critical, and category. Then, data aggregation was carried out to obtain 6140 data. The data can be seen in Table 1.

Table 1: Air Pollution Standard Index (ISPU) Dataset

date_period	date	station	pm10	pm25	so2	co	o3	no2	max	critical	category
202101	1/22/2021	DKI5 (Kebon Jeruk) Jakarta Barat	45	-	21	13	40	15	45	PM10	BAIK
202101	1/23/2021	DKI5 (Kebon Jeruk) Jakarta Barat	80	-	22	44	44	22	80	PM10	SEDANG
202101	1/24/2021	DKI5 (Kebon Jeruk) Jakarta Barat	27	-	14	9	29	-	29	CO	BAIK
...	...	...	...	...	...	...	...	...	...	...	...

2.2. Data Preprocessing

After the data is collected, the cleaning stage is carried out by covering several steps such as changing attribute data types, handling missing or invalid data, and identifying duplicate data. After the preprocessing process is complete, feature selection is carried out to determine the most relevant attributes in the analysis. From a total of 12 attributes available in the Air Pollution Standard Index (ISPU) dataset, 6 attributes are selected that are considered the most important, namely, PM10, PM2.5, SO2, CO, O3, NO2, and attribute categories. The data is then separated into two parts, namely features and targets. The PM10, PM25, SO2, CO, O3, and NO2 attributes are used as features, while the category attribute is used as the target. After that, the data is divided into training data and test data, with a proportion of 80% for training data and 20% for test data.

2.3. Building Model

In this study, the CatBoost Classifier algorithm was used as the main approach to classify the Air Pollution Standard Index (ISPU) categories. CatBoost, developed by Yandex, is one of the Gradient Boosting Decision Tree (GBDT) methods known to be effective in handling categorical features without requiring additional transformations such as one-hot encoding. Another advantage lies in its ordered boosting mechanism, which is designed to minimize overfitting, a common challenge in stepwise learning models.

To understand the internal structure of this algorithm, Figure 2 presents the general architecture of the CatBoost Classifier. The illustration shows how the algorithm forms a model through several layers of learning processes from L1 to L3, which are then compiled into a final model through the prediction stage. Each layer refines the results of previous learning so that the model is able to produce accurate and stable classifications.

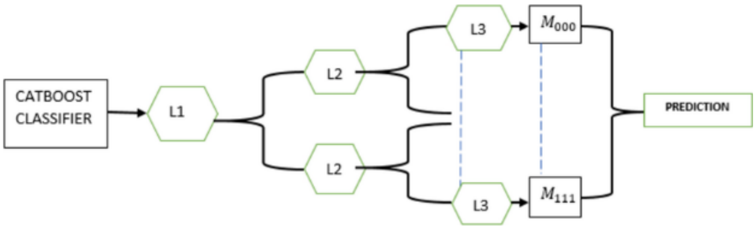

Figure 2: CatBoost Classifier Structure

This image shows the internal workflow of the CatBoost Classifier, starting from the input to the first layer (L1), continuing to the advanced learning layers (L2 and L3), and finally producing predictions through partial models such as. This multi-level mechanism reflects how the algorithm systematically builds and combines decision trees to produce accurate final classifications.

To address class imbalance, class weighting was applied based on the label distribution in the training data. The maximum weight for each class was limited to 100 to prevent extreme bias toward minority classes, particularly the "SANGAT TIDAK SEHAT" category, which only had 3 samples out of a total of more than 5000 data points.

The model developed in this study uses six main air quality parameters as input features, namely PM10, PM2.5, SO2, CO, O3, and NO2. Meanwhile, the target variable consists of four ISPU categories: GOOD, MODERATE, UNHEALTHY, and VERY UNHEALTHY. The model was trained with the following parameter configuration: 300 iterations, a learning rate of 0.1, a tree depth of 6, and a loss function of type 'MultiClass'.

One of the main challenges in model development is the imbalance in class distribution in the target variable. The "SANGAT TIDAK SEHAT" category only has three samples, which is statistically very unrepresentative. Therefore, class weighting is applied to adjust the contribution of each class to the model training process [6]. The maximum weight for each class is limited to a certain value so that the model is not overly biased towards minority classes.

The class weight calculation follows the following formula:

$$\omega_i = \frac{N}{n_i} \tag{1}$$

With:
ω_i: weight of class-i
N : total number of samples in the training data
n_i: number of samples in class-i

Through this approach, classes with less data will be given greater training weight. This strategy aims to improve the model's sensitivity to minority classes without sacrificing classification performance in majority classes.

2.4. Model Evaluation

The model that has been built is evaluated using classification metrics such as accuracy, precision, recall, and F1-score. The evaluation is carried out on test data to measure how well the model classifies each ISPU category. The evaluation results are shown in Table 2, which shows high performance in the majority of categories but weakness in the "SANGAT TIDAK SEHAT" category.

Table 2: Model Performance Evaluation on Testing Data

Category	Precision	Recall	F1-Score	Support
BAIK	0.96	0.93	0.94	101
SEDANG	1.00	0.99	1.00	792
TIDAK SEHAT	1.00	1.00	1.00	135
SANGAT TIDAK SEHAT	0.00	0.00	0.00	3

Based on the above results, the model has a high F1-score in the three main categories (BAIK, SEDANG, and TIDAK SEHAT), indicating that the model can recognize patterns well. However, the zero value in the "SANGAT TIDAK SEHAT" category indicates that the model is unable to predict this category due to data imbalance.

After reviewing the numerical evaluation results in Table 2, it is important to evaluate the prediction distribution visually to gain a more comprehensive understanding of the model's performance. Therefore, a confusion matrix visualization was performed to complement the previous quantitative metric analysis.

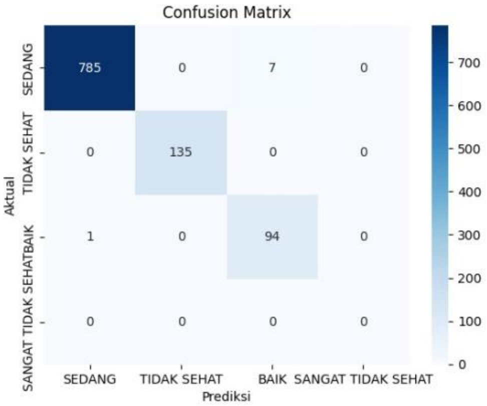

Figure 3: Confusion Matrix of CatBoost Classifier Model Prediction Results

This image shows the confusion matrix of the model's predictions on the test data. The vertical axis shows the actual labels, while the horizontal axis shows the model's predictions. Darker colors indicate higher prediction frequencies. The matrix shows that most predictions are on the main diagonal, indicating that the model has high accuracy in classifying the categories BAIK, SEDANG, and TIDAK SEHAT. The absence of predictions on the "SANGAT TIDAK SEHAT" label underscores the importance of special treatment, such as adding rule-based classification.

To overcome these limitations, a hybrid approach combining machine learning models with rule-based logic was implemented. This approach stipulates that if the PM10 value exceeds 350 or the PM2.5 value exceeds 300, the category will automatically be classified as "SANGAT TIDAK SEHAT."

3. Results and discussion

The CatBoost Classifier model shows excellent classification performance for ISPU categories with sufficient data representation. Based on the evaluation, the average F1-score and accuracy values reached more than 0.98, especially for majority categories such as "SEDANG" and "TIDAK SEHAT". The only weakness appeared in the "SANGAT TIDAK SEHAT" category, which had very limited data.

Uneven data distribution is a major challenge in the classification process. Table 3 shows the distribution of data in each category after preprocessing. It can be seen that the majority of data is in the "SEDANG" category, while "SANGAT TIDAK SEHAT" has only 3 data points, or less than 0.1% of the total.

Table 3: Data Distribution per Category After Preprocessing

Category	Total	Percentage
SEDANG	3992	78.17%
TIDAK SEHAT	649	12.71%
BAIK	463	9.07%
SANGAT TIDAK SEHAT	3	0.06%

As a solution, rule-based logic is used to detect extreme cases that are not well represented by the model. This strategy ensures that classification does not only rely on learning from historical data, but also considers critical domain thresholds for public health.

Additionally, predictions are made per monitoring station to obtain a spatial overview of air quality in the Jakarta metropolitan area. This analysis enables the visualization of ISPU category predictions for the next 30 days based on the last 30 days of data from each station. This approach is relevant for supporting early warning systems and location-based decision-making.

Overall, the developed CatBoost model, combined with a rule-based approach, is capable of providing accurate and responsive predictions of air quality conditions and can be further adapted for broader environmental monitoring systems.

4. Conclusion

This study successfully implemented the CatBoost Classifier algorithm in classifying Air Pollution Index (API) categories based on air quality parameters such as PM10, PM2.5, SO2, CO, O3, and NO2. The evaluation results show that the model has very high performance in classifying the categories "BAIK", "SEDANG", and "TIDAK SEHAT" with an F1-score above 0.94. However, the model is not yet able to predict the "SANGAT TIDAK SEHAT" category due to the very limited and unrepresentative amount of data.

As a mitigation effort for this issue, a hybrid approach with rule-based logic was applied. This approach ensures that extreme cases with high threshold values for PM10 or PM2.5 can still be classified as "SANGAT TIDAK SEHAT," even though they do not appear frequently in historical data.

Station-level predictions also add a spatial dimension to air quality analysis, which is highly beneficial for supporting regional environmental monitoring and early warning systems. Overall, the developed CatBoost model provides accurate and responsive predictions of air quality conditions and can be further adapted for broader environmental monitoring systems.

References

[1] Nababan, A. A., Jannah, M., Aulina, M., & Andrian, D. (2023). Prediksi Kualitas Udara Menggunakan Xgboost Dengan Synthetic Minority Oversampling Technique (Smote) Berdasarkan Indeks Standar Pencemaran Udara (Ispu). JTIK (Jurnal Teknik Informatika Kaputama), 7(1), 214-219.

[2] Kementerian Lingkungan Hidup dan Kehutanan (KLHK). (2020). Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor 14 Tahun 2020 tentang Baku Mutu Udara Ambien. Jakarta: KLHK.

[3] Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). Prediksi kualitas udara menggunakan algoritma K-Nearest Neighbor. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 7(2), 496-507.

[4] Kirono, A. A. H., Asror, I., & Wibowo, Y. F. A. (2022). Klasifikasi Tingkat Kualitas Udara Dki Jakara Dengan Algoritma Naive Bayes. eProceedings of Engineering, 9(3).

[5] Triwibowo, D. N., Ashari, I. A., Sandi, A. S., & Rahman, Y. F. (2021). Enkripsi Pesan Menggunakan Algoritma Linear Congruential Generator (LCG) dan Konversi Kode Morse. *Buletin Ilmiah Sarjana Teknik Elektro*, *3*(3), 194-201.

[6] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449.