JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 7, No. 3. Juni 2025 DOI: https://doi.org/10.34288/jri.v7i3.366
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Shapley Additive Explanations Interpretation of the XGBoost Model in Predicting Air Quality in Jakarta

Adhisa Shilfadianis Iffadah[-1], Trimono[-2], Dwi Arman Prasetya[-n]

Data Science / Faculty of Computer Science
UPN Veteran Jawa Timur
21083010016@student.upnjatim.ac.id, trimono.stat@upnjatim.ac.id, arman.prasetya.sada@upnjatim.ac.id

Abstract

Air quality degradation has become an increasing global problem since 2008, including in Jakarta. By 2024, air pollution in Jakarta is estimated to cause 8,400 deaths and losses of around 34 billion rupiah. To address air pollution, air quality prediction is needed using historical data of Jakarta Air Quality Index from January 2021 to May 2024. The XGBoost ensemble model was chosen for its ability to handle complex data and prevent overfitting. And Shapley Additive Explanations (SHAP) to understand how the model makes decisions. Results showed the XGBoost model achieved MAPE 4.44%. Analysis with Shapley Additive Explanations (SHAP) identified PM2.5 was significantly affected by max and PM10 features, while O3, CO, SO2, and NO2 remained relevant. An increase in PM10 tends to increase PM2.5 concentrations, suggesting the need to control this parameter to improve air quality. These results are important to provide a better understanding of the dynamics of air quality as well as provide a reference for the government in formulating more effective policies or preventive measures in Jakarta.

Keywords: Prediction; Interpretation; XGBoost; Shapley Additive Explanations; Air quality

Abstrak

Penurunan kualitas udara menjadi masalah global yang meningkat sejak 2008, termasuk di Jakarta. Pada 2024, polusi udara di Jakarta diperkirakan menyebabkan 8.400 kematian dan kerugian sekitar 34 miliar rupiah. Untuk menangani polusi udara, diperlukan prediksi kualitas udara menggunakan data historis Indeks Kualitas Udara Jakarta dari Januari 2021 hingga Mei 2024. Model ensemble XGBoost dipilih karena kemampuannya menangani data kompleks dan mencegah overfitting. Dan Shapley Additive Explanations (SHAP) untuk memahami bagaimana model membuat keputusan. Hasil menunjukkan model XGBoost mencapai MAPE 4,44%. Analisis dengan Shapley Additive Explanations (SHAP) mengidentifikasi PM2.5 dipengaruhi signifikan oleh fitur max dan PM10, sementara O3, CO, SO2, dan NO2 tetap relevan. Peningkatan PM10 cenderung meningkatkan konsentrasi PM2.5, menunjukkan perlunya pengendalian parameter ini untuk memperbaiki kualitas udara. Hasil ini penting untuk memberikan pemahaman yang lebih baik tentang dinamika kualitas udara serta memberikan referensi untuk pemerintah dalam merumuskan kebijakan atau langkah preventif yang lebih efektif di Jakarta.

Kata kunci: Prediksi; Interpretasi; XGBoost; Shapley Additive Explanations; Kualitas udara

INTRODUCTION

Air quality refers to the cleanliness of air in a particular area and is commonly measured using the Air Quality Index (AQI)(Agatha 2023). Globally, a decline in air quality has been observed since 2008. In Indonesia, this issue is particularly severe in major urban centers such as Jakarta. According to the Center for Research on Energy and Clean Air (CREA), pollution from coal-fired power plants is a significant contributor to Jakarta's air pollution, resulting in approximately 1,470 deaths annually and economic losses of up to IDR 14.2 trillion (BBC News Indonesia 2023). The IMHE Global Burden of Disease reported more than 123,000 deaths in Indonesia in 2019 due to air pollution(Faqihah Muharroroh Itsnaini 2024). Furthermore, IQAir estimated that in 2024, air pollution in Jakarta caused around 8,400 deaths, with financial losses reaching approximately USD 2.2 million or IDR 34 billion.

As the capital and largest metropolitan city in Indonesia, Jakarta represents the complex dynamics of urban environmental challenges. Air pollution in Jakarta is influenced by various factors, including industrial activity, transportation, and

energy production. Understanding the air quality in Jakarta is not only crucial for protecting public health, but also essential for ensuring sustainable urban development and improving quality of life. Therefore, accurate prediction of air quality and the identification of pollutant contributions are necessary for implementing effective countermeasures.

To address the challenges of predicting air quality in such a complex environment, the use of ensemble machine learning models such as XGBoost has been widely recommended. XGBoost is known for its high predictive performance and built-in regularization techniques, which help mitigate overfitting, especially when dealing with large and complex datasets(Khusna et al. 2023)(Nababan et al. 2023a). However, one major drawback of many machine learning models, including XGBoost, is the lack of interpretability. This limitation poses a challenge in environmental applications, where transparency and interpretability are important for policy-making and public communication.

To overcome this, explainable AI (XAI) approaches such as SHAP (SHapley Additive exPlanations) are increasingly being used. SHAP allows users to understand the contribution of each input feature to the model's predictions, thereby enhancing transparency and trust in the model's output. This study integrates XGBoost with SHAP to provide a more interpretable framework for predicting PM2.5 concentrations in Jakarta.

Several previous studies have demonstrated the effectiveness of XGBoost in predicting PM2.5 levels. For instance, a study in Tianjin, China used hourly PM2.5 data and compared XGBoost with several other models, including Random Forest, SVM, and Multiple Linear Regression. XGBoost achieved the best performance with an RMSE of 17.298, MAE of 11.774, and R^2 of 0.952(Pan 2018). In another study that predicts PM2.5 using daily PM2.5 data and found that XGBoost outperformed other models such as Ridge, Lasso, and AdaBoost, achieving an MAE of 8.27 and RMSE 13.85 (Kothandaraman et al. 2022). In different approach, combined ridge regression with extreme gradient boosting (RR-XGBoost) using air quality sensors and obtained a MAPE with an average of 0.0733 and an average R2 of 0.98(Liu et al. 2021).

Based on the findings from previous studies, it is evident that the XGBoost model performs well in air quality prediction tasks. Nevertheless, the black-box nature of XGBoost limits its interpretability. This study addresses that limitation by combining XGBoost with SHAP to enhance model transparency in the context of PM2.5 prediction in DKI Jakarta.

The primary objective of this research is to develop an interpretable and accurate air quality prediction model using the XGBoost-SHAP framework. This model aims not only to predict PM2.5 concentrations reliably, but also to provide in-depth insights into which features most significantly influence the predictions.

The novelty of this study lies in the application of the XGBoost–SHAP combination for PM2.5 prediction specifically in DKI Jakarta, which has rarely been explored with this level of interpretability. The contribution of this research includes practical insights for public awareness, as well as providing a data-driven reference for policymakers in formulating targeted environmental policies and pollution mitigation strategies.

RESEARCH METHODS

Explanatory Data Analysis (EDA)

This stage is useful to help understand the data which consists of several activities such as understanding the distribution of numerical and categorical data, analyzing correlations between features, and also performing spatial analysis to map the distribution of monitoring stations(Riyantoko et al. 2020).

Data Preprocessing

This stage focuses on data preparation before entering modeling. This stage consists of several activities such as dealing with missing values, performing data transformation such as encoding categorical variables, and feature scaling.

Modelling Extreme Gradient Boosting (XGBoost)

Gradient Boosting method extended with Extreme Gradient Boosting (XGBoost)(Nababan et al. 2023a). The method known as the gradient boosting algorithm is a regression and classification technique that can generate predictive models (Nababan et al. 2023b)(Riyantoko et al. 2021). The goal of this algorithm is to improve processing time efficiency, and includes automation of handling missing data, performing decision trees in parallel, and performing continuous data training to improve accuracy results (Luo et al. 2021). The algorithm itself consists of several weak learning models using loss function model evaluation. The smaller the loss function value, the better the performance of the model(Nababan et al. 2023a).

This algorithm will sum the results of K trees as the final prediction value, which can be written in equation 1(Damaliana, Muhaimin, and Prasetya 2024)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F$$

In equation 1, $\hat{y}_i$ is the predicted value, f_k is the regression result of tree-k, F is the set of decision trees, the following equation is expected to have a value as close as possible to the true value:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

Where $l(y_i, \hat{y}_i)$ is the loss value of the distance between the predicted value and the actual value, some commonly used loss functions are logarithmic loss function, quadratic loss function, and exponential loss function. And Ω which is regularization can be written in the following formula:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \quad (3)$$

In standard terminology, γ serves as a hyperparameter that regulates the complexity of the model, while T represents the number of leaf nodes. λ is the penalty coefficient for leaf weight ω, which is maintained as a constant. These parameters are, λ and ω, determines the complexity of the model and is often determined through empirical means. During the training process, new trees are introduced to accommodate residuals from previous iterations.

XGBoost will perform a Taylor expansion of the objective function, take the first three terms, remove the high order small decimal infinite terms, and change the objective function to:

$$\tilde{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_i(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_k) \quad (4)$$

Where g_i is the first derivative while h_i is the second derivative of the loss function.

The way this XGBoost regressor works is to first initialize the initial model, often in the form of an average of the target variables to be predicted. Then XGBoost will iterate to improve the initial model by gradually adding decision trees to the model. The addition of the decision tree is an attempt to reduce the prediction error generated by the previous trees. For data that is difficult to predict or incorrectly estimated, XGBoost will give greater weight so that subsequent trees focus on correcting these errors. The results of all decision trees used are summed together to produce the final prediction. Each tree contributes to making this final prediction, with trees that are better at predicting the data having a greater influence. XGBoost also applies regularization techniques to avoid overfitting(Kurniawan and Indahyanti 2024). (1)

Model Evaluation

At this stage, a prediction model will be built to predict the value of air quality parameters. In this process, researchers will try to use the XGBoost model. Where both models are machine learning models with different techniques. The XGBoost model will also be hyperparameter tuning using gridsearch with a 10-fold cross validation mechanism. This hyperparameter tuning process is used to obtain the best model parameter scenario so as to get a model with the best performance. When the modeling process is complete, the model will be evaluated using MAPE to determine the accuracy of the prediction model.

One method to calculate the error in the prediction process is the Mean Absolute Percentage Error (MAPE)(Jange 2022)(Astutiningsih, Saputro, and Sutanto 2023). The range of MAPE values is shown in table 1(Maricar 2019)(Trimono, Sonhaji, and Mukhaiyar 2020)(Statistika et al. 2017)

Table 1. MAPE Value Category

Range MAPE	Description
< 10%	Accurate
10% - 20%	Good
20%-50%	Good Enough
>50%	Inaccurate

The MAPE formula is written

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i(x)}{y_i}\right| \quad (7)$$

Where, n is the amount of data, y_i is the true value, and $\hat{y}_i(x)$ is predicted value.

Shapley Additive Explanations (SHAP)

The SHAP approach is a method that allows models to interpret machine learning model predictions that are blackbox or difficult to understand. The goal of SHAP is to explain the prediction of feature x by calculating its contribution to each prediction feature[19]. The SHAP approach uses the formula in equation 3:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\cdot(|p|-|S|-1)!}{|p|!}\cdot[f(S \cup \{i\}) - f(S)] \quad (9)$$

ϕ_i is the shapley value of the feature members against the prediction result. f(S) is the output of a machine learning method that will use a feature set S, and p is the total number of all features. Its final contribution will be defined as the average of its marginal contribution across all possible permutations of the feature set.

Analysis Steps

The following are the analysis steps used in this study:

1. Data collection
2. Exploratory Data Analysis (EDA)
3. Data preprocessing such as handling missing values, encoding data, and dealing with outlier data
4. Modeling the XGBoost algorithm and hyperparameter tunning using Grid Search.
   a. Divide the data into training and testing
   b. Defining parameters for Grid Search
   c. XGBoost model initialization
   d. Grid Search Initialization with Cross-Validation
   e. Applying the best parameter combination
   f. Build XGBoost model with the best parameters Grid Search
   g. Predicting test data using the XGBoost model
   h. Model evaluation using MSE, RMSE, MAPE, and R^2
5. Interpretation using Shapley Additive Explanations (SHAP)
   a. Create a SHAP Explainer to calculate the contribution of each feature to the prediction for each test data sample.
   b. Calculating the SHAP value for each prediction made by the model on the test data
   c. Calculating the average value of the model prediction (expected value)
   Visualization of SHAP values

RESULTS AND DISCUSSION

Dataset

In this study, the authors used secondary data, namely the Air Pollution Standard Index (ISPU) data for DKI Jakarta Province for 2021-2024 which is available on the Satu Data Jakarta website with the website address https://satudata.jakarta.go.id/ . The downloaded data is the Air Pollution Standard Index (ISPU) data in DKI Jakarta Province which contains unitless numbers measured by 5 Air Quality Monitoring Stations (SPKU) in DKI Jakarta Province. This data is data measured every day from January 2021 to May 2024 with a total of 4774 data with numeric and categorical data categories.

Table 2. Information of Dataset

Attributes	Info	Type
stasiun	Location of placement of air monitoring equipment	String
PM10	ISPU measurement result value for particulate matter below 10 microns (PM10)	Integer
PM2.5	ISPU measurement result value for particulate matter below 2.5 microns (PM2.5)	Integer
SO2	ISPU measurement result value for sulfur dioxide parameter	Integer
CO	ISPU measurement result value for carbon monoxide parameter	Integer
O3	ISPU measurement result value for ozone parameter	Integer
NO2	ISPU measurement result value for nitrogen dioxide parameter	Integer
max	the highest value of the ISPU measurement results of several monitored parameters	Integer
critical	the name of the monitored parameter that has the highest ISPU value result	String
category	ISPU measurement result category	String

Exploratory Data Analysis

At this stage, the first activity is to analyze the descriptive statistics of the air quality parameters.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.366
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

	PM10	PM2.5	SO2	CO	O3	NO2
count	4774	4774	4774	4774	4774	4774
mean	49.11	70.97	36.18	12.95	30.57	19.96
min	0	0	0	0	0	0
25%	38	56	25	9	19	12
50%	53	75	38	12	27	18
75%	62	90	50	16	38	26
max	179	287	112	70	181	202
std	19.54	31.39	14.96	6.78	17.76	11.96

Figure 1. Statistic Descriptive

From the descriptive statistics, it is found that all parameters have 4774 data entries, which means there are no missing values. For the PM10 parameter, the average PM10 from January 2021 to May 2024 is 49.11 μg/m^3 which is still within the daily safe limit of PM10. This PM10 has a fairly wide range of values from 0 to 179 μg/m^3. This shows that there are large variations in air quality related to fine particles in the air.

For the PM2.5 parameter, the average PM2.5 from January 2021 to May 2024 is 70.97 μg/m^3 which exceeds the daily safe limit set by the Indonesian government. And the distribution of this parameter shows considerable fluctuation.

The SO2 parameter has an average SO2 from January 2021 to May 2024 is 36.18 μg/m^3 with a range of 0 to 112 μg/m^3. The average for the CO parameter is 12.95 μg/m^3 with maximum value 70 μg/m^3. For the O3 parameter, the average concentration values are 30.57 μg/m^3. While the NO2 parameter has an average NO2 is 19.96 μg/m^3. Next is to see the distribution of DKI Jakarta's air quality categories from January 2021 to May 2024.

Table 3. Distribution of air quality categories

Category	Value
Moderate	3550
Unhealthy	673
Good	505
No Data	42
Very Unhealthy	4

From table 3, it can be seen that from January 2021 to May 2024, air quality in DKI Jakarta is mostly dominated in the moderate and even unhealthy categories. And as many as 4 times reached the category of very unhealthy air quality.

Data Preprocessing

In doing preprocessing, the first thing to do is to look at the empty values. In this data, there are 41 rows of missing values in the critical column. Because the critical column is categorical, then to handle the missing value by using the mode of the critical column, namely PM2.5.

In addition to checking the missing value, because in this research data the NaN value is written as 0. And in this data there are 301 rows of value 0 for PM10 column, 403 rows of value 0 for PM2.5 column, 159 rows of value 0 for SO2 column, 89 rows of value 0 for CO column, 102 rows of value 0 for O3 column, 118 rows of value 0 for NO2 column, 41 rows of value 0 for max column, and 13 rows of value 0 for critical column. For numeric columns such as PM10, PM2.5, SO2, CO, O3, NO2, and Max, handle it using the KNN Imputer method with an approach of 5. KNN Imputer was chosen because it is able to estimate missing values based on the proximity of features between data, thus maintaining the natural structure and patterns between variables in the dataset. This approach is more accurate than simple methods such as mean or median imputation, especially in air quality data that has complex correlations between pollutant parameters. The number of neighbors of K=5 was chosen to achieve a balance between bias and variance in the imputation process. By using five nearest neighbors, imputation becomes quite stable against outliers while remaining sensitive to local patterns in the data. As for the critical column, handle it using the mode of the critical column.

Next is to replace the data to change the Critical column values to make it easier to process, namely by changing PM10 to 1, PM2.5 to 2, SO2 to 3, CO to 4, O3 to 5, and NO2 to 6. Because there are still categorical columns such as the Station and Category columns, data encoding is carried out using the Label Encoder and converting data from strings to integers from the range 0 to 5.

The next preprocessing carried out in this study is to see outlier data and handle it by stamping outliers or winsorizing. In this research data, PM10 has 41 outlier data, PM2.5 has 101 outlier data, SO2 has 2 outlier data, CO has 170 outlier data, O3 has 215 outlier data, and NO2 has 115 outlier data. Stamping outliers or Winsorizing is a method used to deal with outliers by restricting extreme values to within a certain percentile range, so that the influence of outliers on statistical analysis and model performance can be minimized. This technique was chosen because it is able to reduce the distortion caused by extreme values without having to delete data observations, thus maintaining the structure and size of the dataset. Given the characteristics of air quality data that are prone to large fluctuations due to environmental factors.

XGBoost Model Prediction Results

In this study, in predicting air quality using the XGBoost model, Grid Search hyperparameter tuning was also used. With the Grid Search parameters used, namely

Table 4. Grid Search Parameters

Parameter	Value
n_estimators	100, 200
max_depth	3, 5, 7
learning_rate	0.01, 0.1, 0.2

By using negative mean squared error (MSE) evaluation, where negative values are used to maximize the score. And using cross-validation 10 subsets (folds) with 9 subsets as training and 1 subset as testing.

There are 3 types of data division in this study, namely 90% data for testing and 10% data for testing, 80% data for training and 20% data for testing, and 70% data for training and 30% data for testing.

The following are the results of the PM2.5 prediction of the XGBoost model using 10% testing data and with hyperparameter tunning grid search, the best parameters are learning rate 0.1, max_depth 7 and n_estimators 200:

Table 5. PM2.5 prediction result

Date	Actual PM2.5	Predict PM2.5	Difference
2022-06-16	130.5	130.5204	0.020447
2023-11-01	104	103.9786	-0.021355
...	...	...	...
2021-02-10	64	64.5508	0.550812
2021-04-19	92	92.1386	0.138626

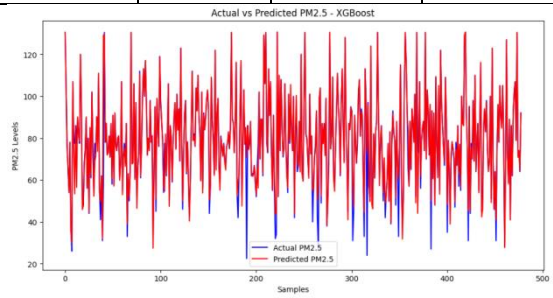

Figure 2. PM2.5 prediction graph

Model Evaluation

In predicting the air quality of DKI Jakarta, a model is not free from errors. Therefore, in making a prediction model, its success in predicting is measured in several ways. Where, in this study using MSE, RMSE, R^2, and MAPE.

Table 6. Evaluation of XGBoost model for PM2.5 parameters

Split Data	MAPE
90:10	4.44%
80:20	4.9%
70:30	4.98%

The XGBoost model in predicting PM2.5 parameters the lowest MAPE is found in 10% testing data, namely 4.44%. In general, the XGBoost model is able to predict PM2.5 parameters accurately, as seen from the MAPE results which are below 10% and the coefficient of determination is close to 100%.

Table 7. Comparison of prediction model performance between this study and several previous studies based on MAPE values

Research	Model	MAPE
Research (Putra et al. 2024)	Hybrid Arima-ANN	13.6%
Research (Fauzan et al. 2024)	LSTM	8.07%
Research (Salsabilla, Fitri Syaharani, and Chamidah 2023)	Ordinary Kriging	19.5%
Current Research	XGBoost	4.44%

When compared to previous studies that predicted PM2.5 with other models, such as Hybrid ARIMA-ANN which resulted in MAPE 13.6%(Putra et al. 2024), LSTM that generates MAPE 8.07%(Fauzan et al. 2024), Ordinary Kriging that generates MAPE 19.5%(Salsabilla et al. 2023). It was found that the XGBoost model had the best performance with a MAPE of 4.44%. This indicates that the XGBoost model is capable of providing more accurate predictions for PM2.5 concentrations compared to traditional statistical methods, hybrid models, and deep learning approaches used in previous studies. The superior performance of XGBoost can be attributed to its ability to capture complex nonlinear relationships and interactions among input features through an efficient gradient boosting framework. Therefore, XGBoost is considered a highly effective model for air quality prediction, particularly for datasets characterized by high variability and nonlinearity, such as PM2.5 measurements.

Interpretation of Shapley Additive Explanations (SHAP)

In this study, the XGBoost model was interpreted using SHAP summary plot to show the contribution of each feature to the prediction of the XGBoost model. Where, the vertical axis contains the features used in the model and is organized by importance or greatest influence. Meanwhile, the horizontal axis shows the influence of the feature on the model prediction. Where, if the SHAP value is positive, it means increasing the prediction and if the SHAP value is negative, it means decreasing the prediction.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i3.366
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

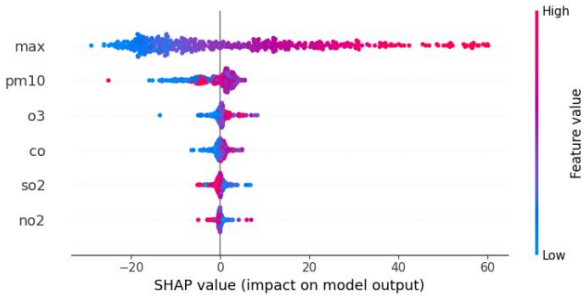

Figure 3. SHAP Summary Plot PM2.5

Figure 3 is a graph of PM2.5 parameters, it is found that max is the feature that has the most influence on prediction seen from the widest spread of SHAP values. And the high value marked by the red plot significantly improves the prediction. PM10 is also a feature that has a large influence that tends to improve predictions. While O3, CO, SO2, and NO2 are lower-level features, they are still relevant. Thus, if there is an increase in the PM10 parameter, it will also result in an increase in the PM2.5 parameter.

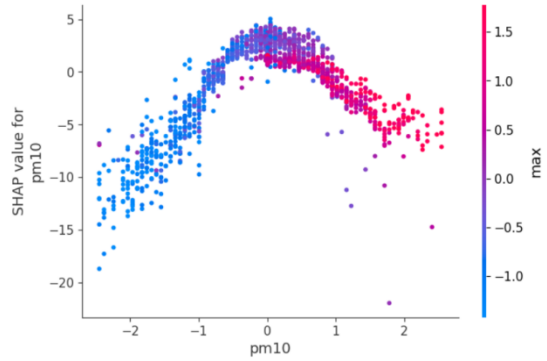

Figure 4. SHAP Dependence Plot Graph

Figure 4 illustrates the relationship between the PM10 feature and the contribution to the PM2.5 model prediction. It is found that PM10 and SHAP Value have a non-linear relationship. At low PM10 values (on the left side), the PM2.5 prediction contribution is negative (low SHAP value). The PM10 value is very high on the right side but decreases again, indicating that there is a saturation effect. While the Max feature has a point with a red color, which means it tends to have a larger SHAP value. This shows that the max feature strengthens the influence of PM10 on PM2.5 prediction. So this graph shows that the effect of PM10 on PM2.5 is strongly influenced by the max feature.

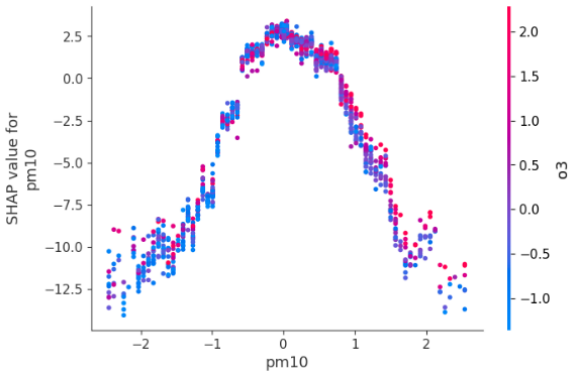

Figure 5. SHAP Interaction Values Graph

Figure 5 is a graph for the SHAP analysis results that illustrates the relationship between PM10 and its contribution to PM2.5. It shows that high O3 values strengthen the influence of PM10 on PM2.5 prediction. Conversely, low O3 values tend to weaken the influence of PM10 on PM2.5. So this graph shows the interaction between PM10 and O3, where the contribution of PM10 to PM2.5 does not only depend on the value of PM10, but is also influenced by the value of O3. To effectively control PM2.5, it is necessary to consider not only PM10 levels but also the interaction effects with other parameters such as O3. These results support the importance of a multidimensional approach in air pollution mitigation strategies.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This research successfully evaluates air quality in DKI Jakarta and predicts PM2.5 concentrations using the XGBoost model which is complemented by SHAP interpretation analysis. The evaluation results of the XGBoost model showed excellent performance in predicting PM2.5, with the lowest MAPE of 4.44%. MAPE values below 10% prove that the model has accurate prediction capabilities.

The SHAP analysis provides important insights into the features that influence PM2.5 prediction. The max feature has the greatest influence on prediction, followed by PM10. Meanwhile, the parameters O3, CO, SO2, and NO2 show a lesser but still relevant influence. The SHAP graph shows that interactions between parameters, such as between PM10 and O3, significantly affect PM2.5 predictions. High O3 values tend to strengthen the influence of PM10 on predictions, while low O3 values weaken it. This confirms the need for a multidimensional approach in air pollution mitigation strategies.

Overall, this study shows that the XGBoost model can be effectively used to predict PM2.5 concentrations with high accuracy. These findings make a significant contribution in understanding the interactions between air quality parameters and provide a basis for decision-making in air pollution mitigation efforts in the DKI Jakarta area.

Suggestion

This study used air quality data from a number of monitoring stations in Jakarta within a certain time period. To improve the generalizability of the model and gain a more comprehensive understanding of air pollution patterns, future researchers are advised to use data with a longer time span and wider area coverage. The use of periodic data that includes seasonal variations, weather conditions, and special events such as forest fires or severe traffic jams can also enrich the analysis results and improve prediction accuracy.

In addition, this study focused on the main pollutant variables commonly found in air quality data. Future researchers are advised to explore additional environmental variables such as air temperature, relative humidity, atmospheric pressure, wind direction and speed, and traffic intensity. The addition of these features can help the model capture more complex relationships between factors that cause air pollution and result in more accurate predictions and more meaningful interpretations.

Finally, although the XGBoost model performed well in this study, future researchers are advised to evaluate alternative models such as LightGBM, CatBoost, and other ensemble models. By conducting a comparison between models, researchers can determine the most optimal approach in terms of prediction accuracy, training efficiency, and interpretability. This evaluation can also strengthen the validity of the findings and make a broader contribution to the development of a reliable and applicable air quality prediction system.

REFERENCES

Agatha. 2023. "Apa Itu Indeks Kualitas Udara (AQI) Dan Bagaimana Cara Menggunakannya?" *Ai Care* .

Astutiningsih, Tiyas, Dewi Retno Sari Saputro, and Sutanto. 2023. "Optimasi Algoritme Xtreme Gradient Boosting (XGBoost) Pada Harga Saham PT. United Tractors Tbk." *SPECTA Journal of Technology* 7(3):632–41. doi: 10.35718/specta.v7i3.1031.

BBC News Indonesia. 2023. "Riset Sebut Polusi Udara PLTU Suralaya Banten 'Menyebabkan 1.470 Nyawa Melayang.'" *BBC*.

Damaliana, Aviolla Terza, Amri Muhaimin, and Dwi Arman Prasetya. 2024. "FORECASTING THE OCCUPANCY RATE OF STAR HOTELS IN BALI USING THE XGBOOST AND SVR METHODS." doi: 10.14710/JSUNIMUS.

Faqihah Muharroroh Itsnaini. 2024. "Kemenkes: Polusi Udara Faktor Resiko Kematian Ke-5 Di Indonesia." *Kompas.Com*.

Fauzan, Fardhi Dzakwan, Dhymas Adhyza Rayhan, Hala Mutiara Putri, and Fitri Kartiasih. 2024. "Peramalan Konsentrasi PM2.5 Menggunakan Model ARCH/GARCH Dan Long Short-Term Memory (Studi Kasus: Kota Jakarta Pusat)." *Infomatek* 26(1):27–44. doi: 10.23969/infomatek.v26i1.12603.

Jange, Beno. 2022. "Prediksi Harga Saham Bank BCA Menggunakan XGBoost." *ARBITRASE: Journal of Economics and Accounting* 3(2):231–37. doi: 10.47065/arbitrase.v3i2.495.

Khusna, Nida Faoziatun, Syifa Aulia, Shinta Amaria, Alfidha Rahmah, Safril Ahmadi Sanmas, and Fatkhurokhman Fauzi. 2023. "Peramalan Kualitas Udara Di Semarang Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA) Forecasting Air Quality in Semarang Using the Autoregressive Integrated Moving Average (ARIMA) Method." *Prosiding Seminar Nasional UNIMUS* 6.

Kothandaraman, D., N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, and Worku Abera. 2022. "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning." *Adsorption Science & Technology* 2022. doi: 10.1155/2022/5086622.

Kurniawan, Wildan, and Uce Indahyanti. 2024. "Prediksi Angka Harapan Hidup Penduduk Menggunakan Metode XGBoost." *Indonesian Journal of Applied Technology* 1(2):18. doi: 10.47134/ijat.v1i2.3045.

Liu, Bing, Xianghua Tan, Yueqiang Jin, Wangwang Yu, and Chaoyang Li. 2021. "Application of RR-XGBoost Combined Model in Data Calibration of Micro Air Quality Detector." *Scientific Reports* 11(1):15662. doi: 10.1038/s41598-021-95027-1.

Luo, Junling, Zhongliang Zhang, Yao Fu, and Feng Rao. 2021. "Time Series Prediction of COVID-19 Transmission in America Using LSTM and XGBoost Algorithms." *Results in Physics* 27. doi: 10.1016/j.rinp.2021.104462.

Maricar, Azman. 2019. "Analisa Perbandingan Nilai Akurasi Moving Average Dan Exponential Smoothing Untuk Sistem Peramalan Pendapatan Pada Perusahaan XYZ." *Jurnal Sistem Dan Informatika*.

Nababan, Adli A., Miftahul Jannah, Mia Aulina, and Dwiki Andrian. 2023a. "PREDIKSI KUALITAS UDARA MENGGUNAKAN XGBOOST DENGAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) BERDASARKAN INDEKS STANDAR PENCEMARAN UDARA (ISPU)." *JTIK*

(Jurnal Teknik Informatika Kaputama) 7(1):214–19. doi: 10.59697/jtik.v7i1.66.

Nababan, Adli A., Miftahul Jannah, Mia Aulina, and Dwiki Andrian. 2023b. "PREDIKSI KUALITAS UDARA MENGGUNAKAN XGBOOST DENGAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) BERDASARKAN INDEKS STANDAR PENCEMARAN UDARA (ISPU)." *JTIK (Jurnal Teknik Informatika Kaputama)* 7(1). doi: 10.59697/jtik.v7i1.66.

Pan, Bingyue. 2018. "Application of XGBoost Algorithm in Hourly PM2.5 Concentration Prediction." in *IOP Conference Series: Earth and Environmental Science*. Vol. 113. Institute of Physics Publishing.

Putra, I. Kadek Pasek Kusuma Adi, Sediono, M. Fariz Fadillah Mardianto, and Elly Pusporani. 2024. "Analisis Prediktif Menggunakan Metode Hybrid Seasonal Autoregressive Integrated Moving Average – Artificial Neural Network Pada Data Konsentrasi PM2.5 Harian Di DKI Jakarta." *G-Tech: Jurnal Teknologi Terapan* 8(1):565–75. doi: 10.33379/gtech.v8i1.3896.

Riyantoko, Prismahardi Aji, Kartika Maulida Hindrayani, Tresna Maulana Fahrudin, and Mohammad Idhom. 2021. *Exploratory Data Analysis and Machine Learning Algorithms to Classifying Stroke Disease*. Vol. 2.

Riyantoko, Prismahardi Aji, Kartika Maulida Hindrayani, Tresna Maulana Fahrudin, and Eristya Maya Safitri. 2020. *Southeast Asia Happiness Report in 2020 Using Exploratory Data Analysis*. Vol. 2.

Salsabilla, Shafira, Amadea Fitri Syaharani, and Nur Chamidah. 2023. "Prediction of PM2.5 in DKI Jakarta Using Ordinary Kriging Method." *Enthusiastic : International Journal of Applied Statistics and Data Science* 48–58. doi: 10.20885/enthusiastic.vol3.iss1.art5.

Statistika, Departemen, Fakultas Sains, Dan Matematika, Universitas Diponegoro, Jl Soedarto, and S. H. Tembalang. 2017. *Valuasi Harga Saham PT Aneka Tambang Tbk Sebagai Peraih IDX Best Blue 2016 TRIMONO, DI ASIH I MARUDDANI*. Vol. 17.

Trimono, Trimono, Abdulah Sonhaji, and Utriweni Mukhaiyar. 2020. "FORECASTING FARMER EXCHANGE RATE IN CENTRAL JAVA PROVINCE USING VECTOR INTEGRATED MOVING AVERAGE." *MEDIA STATISTIKA* 13(2):182–93. doi: 10.14710/medstat.13.2.182-193.