Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 2, Number 2, October 2022, Page 23-29
https://journal.physan.org/index.php/jocpes/index

Prediction Analysis of PM2.5 Concentration Based on Temperature Variables Using XGBoost Algorithm (Case Study: Kemayoran, Central Jakarta)

Valiant Yuvi Syahreza[1], Aviv Maghridlo[1]
[1]State of Meteorology Climatology and Geophysics Agency

Article Info

Article history:

Received September 9, 2022
Revised September 14, 2022
Accepted September 15, 2022

Keywords:

PM 2.5
Temperature
XGBoost
Central Jakarta
Air Quality Prediction

A B S T R A C T

Improvement in air quality in urban areas like Central Jakarta is a big challenge due to high activities of transport, industry, and dense population. This study aims to predict PM2.5 concentrations by utilising the XGBoost algorithm based on temperature data as the main variable. The data was taken from Kemayoran, Central Jakarta, with an observation time span from 01 January 2017 to 12 February 2017. XGBoost was chosen due to the non-linear and complex nature of the data. Based on the results of the test, it shows that the model performance is far from improved, characterized by a high Mean Squared Error (MSE) value and a small R^2 score. These performance limitations are driven by the small amount of data and the absence of other supporting variables such as air humidity, wind speed, and rainfall. The high PM2.5 concentration was contributed by the research location in Kemayoran, one of the most densely populated areas with high industrial activity and fossil-fuelled transport. This study provides evidence to support the addition of supporting variables and the extension of the observation time span to enhance model accuracy. Therefore, the XGBoost algorithm can be used as a promising solution for air quality prediction in urban cities where air pollution has reached its peak.

Corresponden Author:

Valiant Yuvi Syahreza,
State of Meteorology Climatology and Geophysics Agency
Tangerang City, Banten, Indonesia
Email: backupsementara30@gmail.com

1. INTRODUCTION

As it is one of the most populated cities in Indonesia, Central Jakarta has been facing serious problems in terms of air quality [1]. High levels of air pollution are caused by transport areas, urbanisation patterns, and geographical features [2]. One of the main components of air pollution that has a direct adverse impact on public health is particulates (PM 2.5). PM2.5 can penetrate the human respiratory system and cause respiratory tract irritation and chronic diseases.

In recent years, machine learning approaches have become a promising solution for modelling complex relationships [3]. One of the widely used algorithms is Extreme Gradient Boosting (XGBoost), which is known for its ability to handle data with non-linear relationships and capture patterns that are difficult to detect by traditional methods [4].

However, studies in the Jakarta area, especially in Central Jakarta, are scarce. Most studies are limited to static or linear regression analyses, which often fail to capture the complex interactions between environmental factors and PM2.5 concentrations [5]. This gap is what this study attempts to fill by using a more advanced machine learning algorithm, namely XGBoost, to analyse and predict PM2.5 concentrations using Central Jakarta temperature data [4].

This research aims to develop a prediction model for PM2.5 concentrations based on temperature data using the XGBoost algorithm [6]. This research uses data from OneAQ for PM2.5 concentration and BMKG for temperature data in the Central Jakarta area [7]. The main contributions of this research are:

- Determine the relationship pattern between air temperature and PM2.5 concentration in Central Jakarta.
- Develop a machine learning-based prediction model using XGBoost to model the relationship.
- Evaluate model performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2).
- Provide insights into the limitations of the data and model, as well as recommendations for future research.

This research is expected to contribute to the development of a more accurate air quality prediction system, so that it can support air pollution mitigation efforts in urban areas.

2. RESEARCH METHOD

To analyse the relationship between air temperature and PM2.5 concentration in the Central Jakarta area, this study uses machine learning methods [8]. To generate the prediction model, the algorithm used is XGBoost, which has been selected for its reliability in modelling non-linear patterns and proven performance in several prediction studies [9]. The data used includes PM2.5 concentrations from the OneAQ platform and temperature data from the BMKG database [10].

The methodological process used in this research is as follows:

A. Data Collection

To analyse the data, this study used two main sources. Data on PM2.5 concentrations were taken from the OneAQ platform, which provides real-time air quality measurements. This information includes daily PM2.5 concentrations in units of micrograms per cubic metre (μg/m3) in the Central Jakarta area. For now, information on air temperature is collected from the BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) database [11][12]. This includes daily average data in degrees Celsius. These two data sources were deemed suitable for conducting analyses on how air quality in the area correlates with meteorological parameters. The data was taken with a time span from 1 January 2017 to 12 February 2017.

B. Data Processing

To ensure that the data from both sources could be used effectively in the analyses, processing steps were undertaken [13]. The first step was data cleansing. This means that empty data or invalid values such as -999 are removed. To ensure efficiency and flexibility in handling various data formats, this cleaning process was performed using Python. Next, an interpolation method was used to fill the data gaps from the 14th to the 16th. Simple linear interpolation was used to fill these gaps, which can be calculated using the following formula [14]:

$$y = y_0 + \frac{(x - x_0)}{(x_1 - x_0)}(y_1 - y_0) \quad (1)$$

Where y_0 and y_1 are the known data values at points x0 and x1 and x is the missing data point.
After interpolation, data from both sources were combined in a uniform time format (datetime) with daily resolution. In addition, data transformation is performed to add new features such as year, month, and day columns, which allow for the analysis of seasonal patterns. If required, normalisation is used to scale the variables and improve the performance of the machine learning model.

C. Model Evaluation

The performance of the XGBoost model is evaluated using statistical metrics, such as Mean Squared Error (MSE) to measure the average squared error between predicted and actual values, and R-squared (R^2) to assess the extent to which the model is able to explain data variability[15]. In addition to quantitative evaluation, the predicted results are compared with the actual data through scatter plot visualisation[16]. This visualisation helps to understand the fit of the model to the actual data and gives an idea of the accuracy of the predictions[17].

D. Corelation Analysis

To evaluate the performance of the XGBoost model, statistical metrics such as Mean Squared Error (MSE) are used to measure the average squared error between predicted and actual values, and R-squared (R2). In addition to quantitative evaluation, the predicted results are compared with the actual data through scatter plot visualisation. This visualisation helps to understand the fit of the model to the actual data and gives an idea of the level of fit.

E. Limitation of the Study

There are several limitations in this study that need to be noted. First and foremost, the temperature data used only looked at average temperature variables and did not include environmental factors such as wind speed, humidity or air pressure, all of which can affect PM2.5 concentrations [18]. Secondly, as the PM2.5 data from OneAQ is only a localised measurement, it may not reflect spatial variations across the whole of Central Jakarta. Lastly, the amount of data available is very limited, which may affect the reliability [19]. However, the results of this study are expected to provide important information on how air temperature and air quality in central Jakarta correlate with each other [20]. This can be achieved due to the systematic use of the research methodology.

3. RESULT AND DISCUSSION

This study uses the XGBoost model to analyse the relationship between air temperature and PM2.5 concentration in Central Jakarta. To understand the data patterns, various visualisations were used to assess the model performance and compare the predicted and actual values.

A. Data Collection

To explore the relationship between air temperature and PM2.5 concentration, a correlation analysis was conducted, which was also visualised as a scatter plot. The scatter plot results show the variation between temperature and air pollution levels. The correlation value between temperature and PM2.5 was calculated using the Pearson correlation coefficient. Based on the results of the analysis, a correlation value was obtained that illustrates how much the relationship between these two variables is.

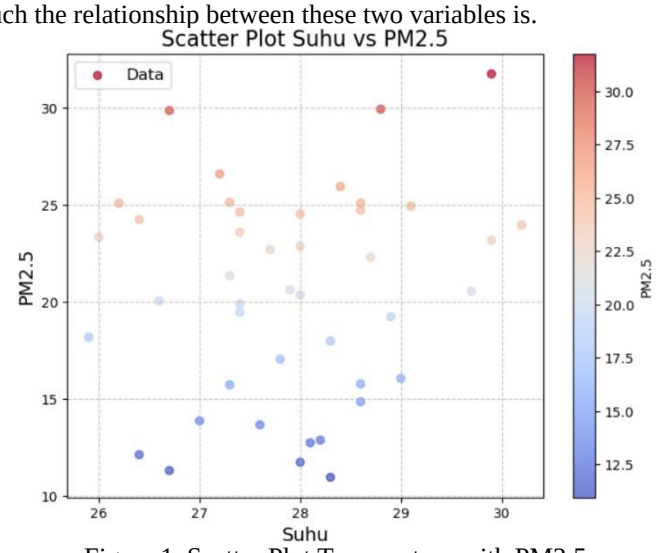

Figure 1. Scatter Plot Temperature with PM2.5

Above is the scatter plot graph showing the relationship between temperature and PM2.5 concentration by the nature of the data analyzed. The horizontal axis (x-axis) represents the temperature, ranging from 26 to 30 degrees Celsius, while the vertical axis (y-axis) represents the PM2.5 concentration, ranging from 10 to 30 µg/m³. This data is visualized in color, where bluer colors represent low PM2.5 concentrations and redder colors represent high PM2.5 concentrations.

This plot shows that there is large variation in PM2.5 values at lower temperatures. Values below 28 degrees Celsius have a wide distribution of low to high PM2.5 concentrations, while above 29 degree Celsius, PM2.5 values seem more concentrated around a smaller area and tend to show higher values towards 25 to 30 µg/m³.

This pattern gives an indication of the relationship between temperature and PM2.5 distribution, though there is not any clear linear relationship from it. The colors used in this plot help identify that PM2.5 values vary not only due to temperature but may also be related to other environmental factors which might not be directly visible from this graph.

From this visualization, it can also be noticed that data points with high PM2.5 values (red color) are less frequent than those with moderate values (orange to blue color). The diverse distribution of data shows that PM2.5 concentrations are not only dependent on temperature but also other factors such as atmospheric conditions, local pollution sources, or temporal factors like seasonality.

It can also be seen from this graph that the color scheme provides further details into the intensity of PM2.5 at various temperatures. For instance, in the ambient temperature range of 27 to 28 degrees Celsius,

points with variable colors can be seen. These would then mean that such ambient temperatures can support variable amounts of PM2.5 build-up in the air under specific conditions.

B. Temperature and PM2.5 Distribution

To further understand the pattern of distribution of temperature variables and PM2.5 concentrations, a probability distribution analysis was done with visualization in Kernel Density Estimation. This is useful in showing the distribution of data without assuming any underlying distribution, such as a normal distribution. The resultant graphs show information about the density of data over a range of values for both temperature and PM2.5.

As shown in Figure 2, the analysis results show different distributions between the two variables.

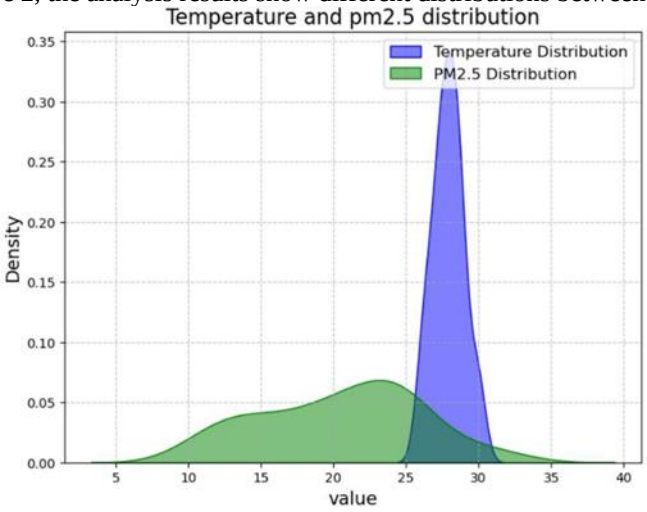

Figure 2. Scatter Plot Temperature with PM2.5

This graph below shows that the temperature is centered between 27 and 30 degrees Celsius. The highest density of the temperature data is at about 29 degrees Celsius, which means the majority of the temperature observations are to be found within this value. This is reflected in the blue coloured graph that shows a very sharp peak in distribution, therefore indicating relatively low variation of temperature around its peak value.

In contrast, the distribution of PM2.5 concentration, represented by the green graph, is more scattered compared to temperature. PM2.5 concentration exhibits a much lower density peak, at around 20 µg/m³, with a long distribution tail up to higher values. This reflects that PM2.5 is more variable than temperature, probably due to a wide variety of environmental factors such as human activities, pollution sources, or other meteorological conditions.

The above difference in the distribution characteristics means that though temperature may have an influence on PM2.5, higher variability of PM2.5 than temperature indicates other factors affecting this pollutant concentration.

C. XGBoost Prediction Output

Comparing predicted results with actual values to evaluate model performance is an important step in the data analysis process. The aim is to find out how close the model's predictions are to the actual data. Scatter plot, which is a commonly used visualisation technique, can show the accuracy and pattern of the relationship between the two variables through the distribution of points on the graph. In addition, to evaluate the accuracy of the model, evaluation metrics such as R2 Score and Mean Squared Error (MSE) are often used. The process of analysing PM2.5 prediction models will be easier to understand with the help of these graphs.

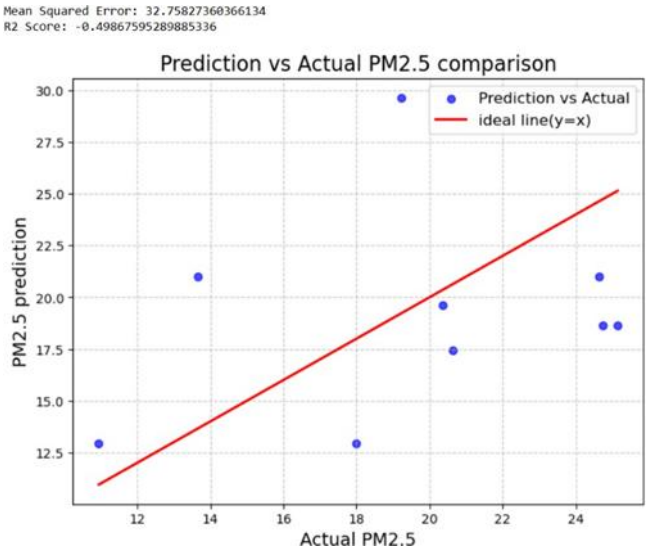

Figure 3. Comparison between Prediction and Actual of PM2.5

Above is the scatter plot graph comparing the predicted PM2.5 values on the Y-axis against the actual PM2.5 values on the X-axis. Each blue dot in the graph represents a pair of actual and predicted values, where the spread of the dots indicates the extent to which the predicted results are close to the actual values. Ideally, all points would lie exactly on the red line, which is the ideal line with the equation ????=????, if the prediction has high accuracy. However, in this graph, most of the points are not near the ideal line, which indicates a large difference or error between the predicted and actual values.

The red line in the graph serves as the ideal reference with which every perfect prediction shall be parallel to the actual value. However, points which are scattered far from the line indicate that the prediction model has been unable to provide accurate results. This is supported by the Mean Squared Error value of 32.76, showing that the average square of the difference between actual and predicted values is huge. This value is indicative of the fact that this model has great deviations in its predictions; the greater the MSE, the worse the performance of the model.

Moreover, the obtained value of R2 Score is -0.49, showing extremely poor model performance. The negative R2 Score value means that this prediction model is worse than using the average of actual values as a prediction. Theoretically, the best value for R2 Score is 1, indicating perfect prediction, while a value of 0 indicates that the prediction is no better than the average. These are negative values; this itself says that the model is unable to capture the pattern present in the data.

The scatter plot visually distributes the points in an irregular pattern, without following the line of ideality. Those above the line indicate overestimation by the model on the actual value, and below the line shows the vice-versa scenario. This deviation reflects that the prediction has a large and inconsistent error. Therefore, it can be concluded from this graph that the performance of the model in predicting PM2.5 values is far from satisfactory and needs further review for improving the accuracy of the prediction

4. CONCLUSION

It is deduced from the experimental result by using the XGBoost algorithm that this model might be a solution to predict the concentration of PM2.5, but the result derived from it is still far from optimal. It shows an extremely high value for the Mean Squared Error (MSE) with an R2 score similarly as low. One of the primary reasons for low prediction accuracy has to do with limited data on which the research is based, whereas it only covers the period from 01 January 2017 to 12 February 2017. This small length of time causes the inability of the model to learn more complex and seasonal variation patterns.

Moreover, other external factors might have influenced the results of the prediction: for example, the absence of further parameters within the analysis. Air humidity, temperature, wind speed, or rainfall are some of the important factors that greatly influence the PM2.5 concentration and have not been considered in the model. The geographical location where the data collection is done also plays an important role, whether it is in an urban or rural area. Urban areas, with high population density, industrial activity, and a high number of fossil-fuelled vehicles, tend to have higher pollution compared to rural areas. This needs further study in order to understand the contribution of the environment to variations in PM2.5 concentrations.

Some steps for improving the performance of the XGBoost model: increasing the amount of data by extending the observation time span so that the seasonal patterns and long-term trends are captured. Adding supporting parameters, such as air humidity, temperature, wind speed, rainfall, and industrial and transport

activity data to provide a more comprehensive context for the analysis. Thirdly, the study of geographical location will be done in order to understand the characteristics of the environment where the data is collected and the influence of industrial or transport activities. Fourth, hyperparameter tuning for better performance: learning rate, max depth, and n_estimators.

With these steps, it is expected that the XGBoost model can provide more accurate prediction results, have a higher correlation value with actual data, and can be used as a reliable solution in air quality analysis.

REFERENCE

[1] A. Assayuti, Y. Pujowati, A. Abeng, and D. Kamal, "Impact of air Pollution, Population Density, Land Use, and Transportation on Public Health in Jakarta," J. Geosains West Sci., vol. 1, pp. 35–43, 2023, doi: 10.58812/jgws.v1i02.391.

[2] B. Haryanto, "Climate Change and Urban Air Pollution Health Impacts in Indonesia," in Climate Change and Air Pollution: The Impact on Human Health in Developed and Developing Countries, R. Akhtar and C. Palagiano, Eds., Cham: Springer International Publishing, 2018, pp. 215–239. doi: 10.1007/978-3-319-61346-8_14.

[3] A. Masood et al., "Improving PM2.5 prediction in New Delhi using a hybrid extreme learning machine coupled with snake optimization algorithm," Sci. Rep., vol. 13, no. 1, pp. 1–17, 2023, doi: 10.1038/s41598-023-47492-z.

[4] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai," Aerosol Air Qual. Res., vol. 20, no. 1, pp. 128–138, 2020, doi: 10.4209/aaqr.2019.08.0408.

[5] A. X. V. I. Simp and S. Remoto, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data Mehdi," no. 1992, pp. 6425–6432, 2013.

[6] Q. Yang, Q. Yuan, T. Li, H. Shen, and L. Zhang, "The relationships between PM2.5 and meteorological factors in China: Seasonal and regional variations," Int. J. Environ. Res. Public Health, vol. 14, no. 12, 2017, doi: 10.3390/ijerph14121510.

[7] P. Zhan et al., "Recent abnormal hydrologic behavior of Tibetan lakes observed by multi- mission altimeters," Remote Sens., vol. 12, no. 18, 2020, doi: 10.3390/RS12182986.

[8] T. Wang et al., "Secondary aerosol formation and its linkage with synoptic conditions during winter haze pollution over eastern China," Sci. Total Environ., vol. 730, p. 138888, 2020, doi: https://doi.org/10.1016/j.scitotenv.2020.138888.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[10] C. Jin, Y. Wang, T. Li, and Q. Yuan, "Global validation and hybrid calibration of CAMS and MERRA-2 PM2.5 reanalysis products based on OpenAQ platform," Atmos. Environ., vol. 274, p. 118972, 2022, doi: 10.1016/j.atmosenv.2022.118972.

[11] J. Guo et al., "Impact of diurnal variability and meteorological factors on the PM2.5 - AOD relationship: Implications for PM2.5 remote sensing," Environ. Pollut., vol. 221, pp. 94–104, 2017, doi: https://doi.org/10.1016/j.envpol.2016.11.043.

[12] C.-H. Wu, I.-C. Tsai, P.-C. Tsai, and Y.-S. Tung, "Large–scale seasonal control of air quality in Taiwan," Atmos. Environ., vol. 214, p. 116868, 2019, doi: https://doi.org/10.1016/j.atmosenv.2019.116868.

[13] G. Shreya, B. Tharun Reddy, and V. S. G. N. Raju, "Air Quality Prediction Using Machine Learning Algorithms," Lect. Notes Networks Syst., vol. 840, no. 2, pp. 465–473, 2024, doi: 10.1007/978-981-99-8451-0_39.

[14] J. Zhou and Z. Huang, "Recover Missing Sensor Data with Iterative Imputing Network," CoRR, vol. abs/1711.07878, 2017, [Online]. Available: http://arxiv.org/abs/1711.07878

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016, pp. 785– 794. doi: 10.1145/2939672.2939785.

[16] Z. Ali and A. Burhan, "Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model," Asian J. Civ. Eng., vol. 24, pp. 1–16, 2023, doi: 10.1007/s42107-023-00651-z.

[17] G. Shmueli and O. Koppius, "Predictive Analytics in Information Systems Research," MIS Q., vol. 35, pp. 553–572, 2011, doi: 10.2139/ssrn.1606674.

[18] H. Zheng et al., “Achievements and challenges in improving air quality in China: Analysis of the long-term trends from 2014 to 2022,” Environ. Int., vol. 183, p. 108361, 2024, doi: https://doi.org/10.1016/j.envint.2023.108361.
[19] M. Diao et al., “Methods, availability, and applications of PM(2.5) exposure estimates derived from ground measurements, satellite, and atmospheric models.,” J. Air Waste Manag. Assoc., vol. 69, no. 12, pp. 1391–1414, Dec. 2019, doi: 10.1080/10962247.2019.1668498.
[20] Y. Zhang, S. X. Chen, and L. Bao, “Air pollution estimation under air stagnation—A case study of Beijing,” Environmetrics, vol. 34, no. 6, p. e2819, 2023, doi: https://doi.org/10.1002/env.2819.