Contents list available at CBIORE journal website

Journal of Emerging Science and Engineering

Journal homepage: https://journal.cbiore.id/index.php/jese/index

Research Article

Hyperparameter optimization for hourly PM2.5 pollutant prediction

Aziz Jihadian Barid[a*] and H. Hadiyanto[b]

[a] *Master Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia*

[b] *Chemical Engineering Department, Faculty of Engineering, Diponegoro University, Semarang, Indonesia*

Abstract. Air pollution, particularly the presence of Particulate Matter (PM) 2.5, poses significant health risks to humans, with industrial growth and urban vehicle emissions being major contributors. This study utilizes machine learning techniques, specifically K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms, to predict PM2.5 levels. A dataset from Kaggle consisting of PM2.5 and other pollutant parameters is preprocessed and split into training and testing sets. The models are trained, evaluated, and compared using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics. Additionally, hyperparameters are applied to optimize the models. Results show that SVM with hyperparameters performs better, indicating its potential for accurate air quality prediction. These findings can aid policymakers in implementing effective pollution control strategies.

Keywords: PM2.5, K-Nearest Neighbor, Support Vector Machine, Hyperparameter

Received: 2nd February 2024; Revised: 17th March 2024; Accepted: and April 2024; Available online: 8th April 2024

1. Introduction

Air is a natural resource that is very important for the survival of humans and other living creatures. The quality of air greatly influences human health, because air is an important component and is a basic need for living creatures at all times. However, industrial growth and the mobility of motorized vehicles in big cities are the main causes of air becoming polluted, one of which is Particulate Matter (PM) 2.5, which is a fraction of air pollutants, which has become an important indicator in environmental management, especially human health (Pak et al., 2020)

The presence of dust particulates, if it is still within the limits specified in the regulations, is still considered a safe condition for the body. However, if the concentration of dust particulates is more than the threshold value, then this can cause problems with human health. These health problems include respiratory disorders, decreased lung function, allergies and bronchitis (Sinolungan et al., 2009) Because air pollution is the main problem in daily life that often occurs, these problems must be resolved immediately and properly. preventive measures must also be taken.

According to the IQAir 2021 world air quality report, Indonesia is ranked 17th with the highest concentration of PM2.5 (Kennial Laia, 2022). In such conditions, experts can utilize computer technology, machine learning, data mining and other tools to collect precise data from monitoring stations. This data can help address pollution and estimate air quality. One technology that can help in this classification is machine learning.

Research has been carried out to predict air pollutants using machine learning by adopting multi-task learning to predict air quality, with prediction of air quality values as the main task (regression task) and prediction of air quality levels as an additional task (classification task).

Mobile monitoring of traffic-related particulate matter (TRPM, including PM10, PM2.5, and BC) and CO2. A random forest land use regression (RF-LUR) model for TRPM and CO2 was developed to quantify the contribution of relevant influential factors. Finally, the accuracy of the LUR and RF-LUR prediction models is evaluated and compared. The research results show that pollutants in environmental scale areas have significant spatial-temporal heterogeneity. Our results show that RF-LUR can be effectively applied to neighborhood-scale traffic pollutant prediction.

A number of studies have been carried out on this matter using various regression techniques such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Decision Trees which have all proven successful in predicting air quality (Umri et al., 2021). Even though the prediction results are reliable, there is a tendency for the model to experience overfitting (Biancofiore et al., 2017), one way to overcome this is to use Hyperparameters in the prediction process.

Considering the serious potential of this air quality problem, this research aims to develop and validate an innovative PM2.5 prediction model by utilizing the computational capabilities of machine learning using KNN and SVM (Hyperparameter). Through a comprehensive evaluation of the model's performance and its relevance to local conditions, this study is expected to provide important insights for policymakers to adopt effective and sustainable intervention strategies in combating air pollution.

* Corresponding author
Email: azizjb100@gmail.com (A.J.Barid)

https://doi.org/10.61435/jese.2024.e15
ISSN: p-ISSN: 3026-0817; e-ISSN: 3026-0183.The Author(s). Published by CBIORE

2. Method

The methodology developed to improve the previous results. aims at classifying air quality according to the Air Quality Index (AQI). Data preprocessing and extraction are required before training the prediction model. The air quality prediction process used in this study is illustrated (Fig 1).

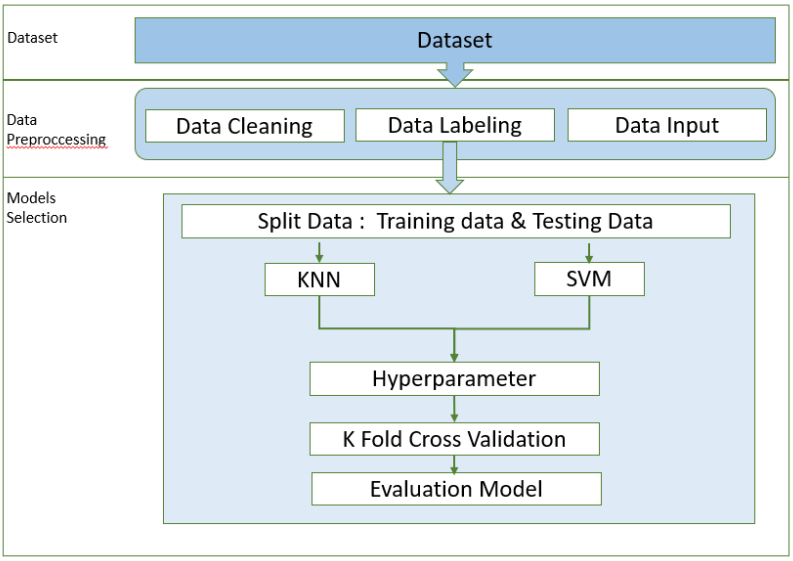


Figure 1. research methodology

2.1 Air Quality Dataset

The research began with collecting datasets obtained from the Kaggle repository in .xlsx format. This dataset consists of around 8000 rows of data with 6 parameters: pm10, pm2.5, so2, co, o3, and no3, but only PM2.5 is used for prediction.

2.2 Labeling & Processing

Before the model implementation process, the dataset undergoes a preprocessing stage. This stage includes deleting irrelevant data, cleaning empty data, identifying and deleting outlier data and taking the previous 5 rows to predict the next 1 row. The cleaned data then undergoes transformation for ML modeling using Scikit-learn (Pedregosa FABIANPEDREGOSA et al., 2011).

2.3 Split Data

The next stage includes dividing the dataset into a training set and a testing set in the proportion of 80:20. The model is trained using the data in the training set, and the model performance is evaluated using the test set. Data sharing with time series models to meet specific training and testing data needs. Each data point serves as input in the research procedure, used for algorithmic training and testing.

2.4 Model Implementation

KNN algorithm is based on the distance between data points. It tests data by calculating the distance to training data and selects the k-nearest neighbors to classify the data. The algorithm considers the k data samples closest to the test sample and associates the majority class with the test sample (Arora et al., 2023)

$$d_i = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2 + (X_i - Y_i)^2} \qquad (1)$$

SVM stands out as one of the most widely employed supervised learning algorithms, extensively utilized for addressing both classification and regression problems. Nevertheless, its primary application in the realm of machine learning is focused on classification tasks (Akhter & Miller, 2023)

2.5 Hyperparameter

After pre-processing and data splitting steps, the data must be trained with SVM & KNN algorithms to build the model. This section provides a brief overview of machine learning algorithms and then applies hyperparameters.

2.6 K-Fold Cross Validation

Cross validation is an additional method of data mining techniques that aims to obtain maximum accuracy. This method is often referred to as k-fold cross validation where k times are tried for one model with the same parameters. (Budi Santosa & Ardian Umam, 2018)

3. Results and Dicsussion

In this research, there are approximately 8527 datasets presented in Table 2, divided into an 80:20 ratio, where 80% is used as training data and 20% as testing data. Table 1 is the PM2.5 pollutant value on 15/10/2017 from 04.00.00 to 13.00.00.

Table 1
Dataset Pollutant PM2.5

Tanggal	Jam	PM2.5
10/15/2021	04:00:00	12
10/15/2021	05:00:00	12
10/15/2021	06:00:00	13
10/15/2021	07:00:00	13
10/15/2021	08:00:00	13
10/15/2021	09:00:00	13
10/15/2021	10:00:00	13
10/15/2021	11:00:00	13
10/15/2021	12:00:00	14
10/15/2021	13:00:00	14

In this study, 5 previous lines were taken to predict 1 next line as shown in Table 2

Table 2

var_1	var_2	var_3	var_4	var_5	target
12	12	13	12	12	12
12	13	12	12	12	12
13	12	12	12	12	13
12	12	12	12	13	13
12	12	12	13	13	13
12	12	13	13	13	13
12	13	13	13	13	13
13	13	13	13	13	13
13	13	13	13	13	14
13	13	13	13	14	14

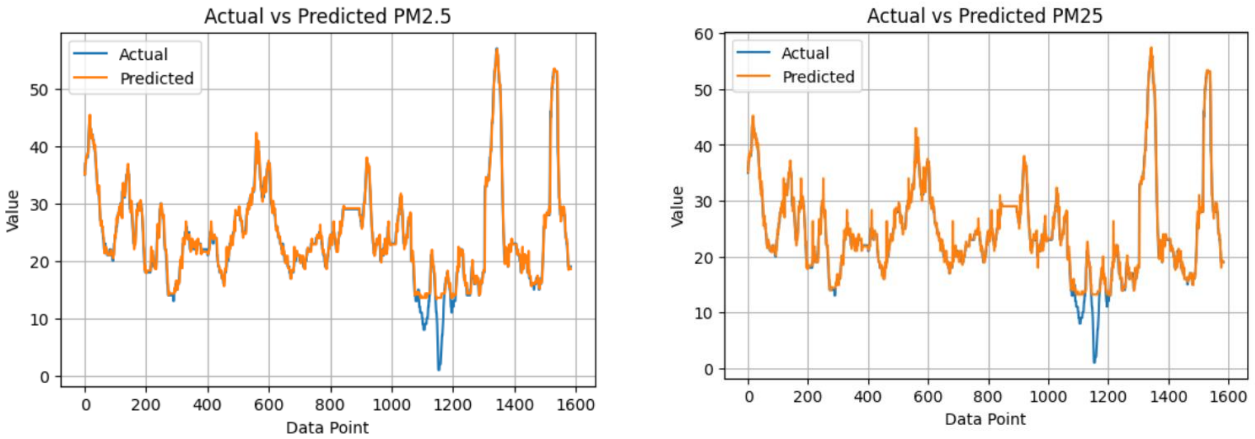
Table 2. contains columns var_1, var_2, var_3, var_4, var_5 is the input value or X and target is the output or prediction

3.1 Model evaluation

Model evaluation for each algorithm used in the research was carried out using the MSE and RMSE matrices to determine the error level in each model. Validation provides a comprehensive evaluation of overall performance through 5-fold cross-validation, measured in MSE and RMSE Matrices. There is a comparison of MSE and RMSE between the k-nearest neighbors algorithm and a support vector machine which has been optimized using hyperparameters so as to produce increased MSE results with a smaller error rate in the SVM + hyperparameter algorithm as seen in Table 3 and visualized in Figure 2 KNN and Figure 3 SVM. The attached graph depicts model accuracy from test and prediction data.

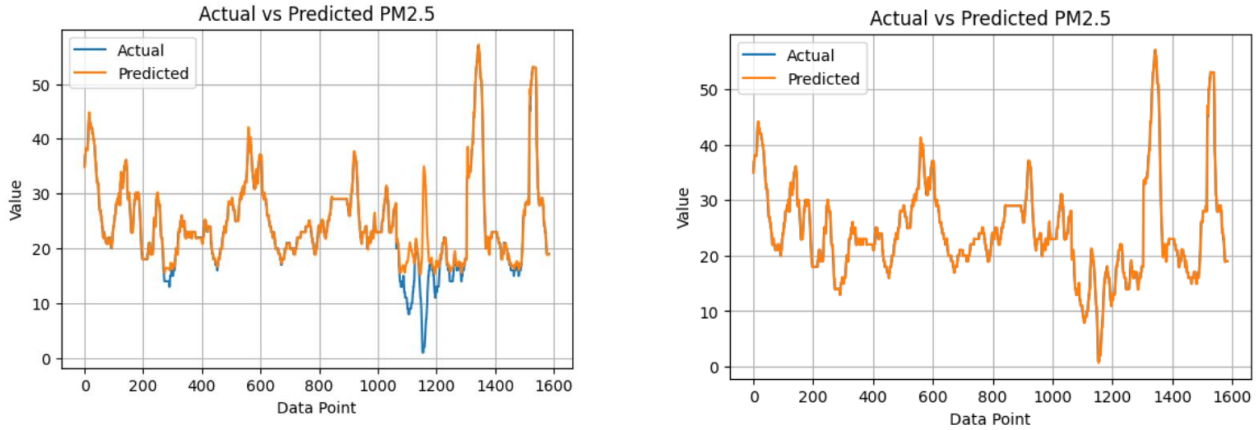
Tabel 3
Comparation Evaluated model

No	Algoritm	MSE	RMSE
1	K-Nearest Neigbor	2.25	1.59
2	Support Vector Machine	10.86	3.29
3	K-Nearest Neigbor + Hyperparameter	2.56	1.60
4	Support Vector Machine + Hyperparameter	0.71	0.84



(a) (b)

Figure 2. Evaluate KNN (a) Before Hyperparameter, (b) After Hyperparameter



(a) (b)

Figure 3. Evaluate SVM (a) Before Hyperparameter, (b) After Hyperparameter

Table 3, Figure 2 and Figure 3 show that the use of hyperparameters in both models has increased performance, but there is a significant increase in the SVM model. The importance of proper data preprocessing and determining the right parameters can improve algorithm performance.

4. Conclusion

This paper uses machine learning techniques to predict PM2.5. Researchers also utilized datasets collected from Kaggle. This research uses the KNN and SVM algorithms which are optimized using hyperparameters. The algorithm is implemented using Python to achieve the best results that can provide the best performance. The main objective of this research is to test the optimal aspects of machine learning algorithms targeted for PM2.5 pollutant prediction. Experimental results show that the performance of SVM using hyperparameters is better with a lower error rate. Therefore, this research can be a reference and provide accurate information for fast and responsive decision making in air pollution management. For method development, emphasis can be placed on adding a segmentation step in the pre-processing process, with a focus on air quality features. The findings from this research can be used as a basis for developing a more efficient and accurate air quality monitoring system, thereby making a positive contribution to decision making regarding air quality management and improvement.

References

Akhter, S., & Miller, J. H. (2023). BaPreS: a software tool for predicting bacteriocins using an optimal set of features. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05330-z

Arora, P., Periwal, N., Goyal, Y., Sood, V., & Kaur, B. (2023). iIL13Pred: improved prediction of IL-13 inducing peptides using popular machine learning classifiers. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05248-6

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., & Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research*, *8*(4), 652–659. https://doi.org/10.1016/j.apr.2016.12.014

Budi Santosa, & Ardian Umam. (2018). *Data Mining Dan Big Data Analytics : Teori dan implementasi mengunakan Python & Apache Spark* (2nd ed.). Penebar Media Pustaka.

Kennial Laia. (2022, March). *Laporan IQAir: Indonesia Peringkat ke-17 Negara Paling Berpolusi*. Https://Betahita.Id/News/Detail/7310/Laporan-Iqair-Indonesia-Peringkat-Ke-17-Negara-Paling-Berpolusi.Html.Html.

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment*, *699*. https://doi.org/10.1016/j.scitotenv.2019.07.367

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Sinolungan, J. S. V, Psikologi, B., Kedokteran, F., Sam, U., & Manado, R. (n.d.). *DAMPAK POLUSI PARTIKEL DEBU DAN GAS KENDARAAN BERMOTOR PADA VOLUME DAN KAPASITAS PARU*.

Umri, S. S. A., Firdaus, M. S., & Primajaya, A. (2021). ANALISIS DAN KOMPARASI ALGORITMA KLASIFIKASI DALAM INDEKS PENCEMARAN UDARA DI DKI JAKARTA. *Jurnal Informatika Dan Komputer) Akreditasi KEMENRISTEKDIKTI*, *4*(2). https://doi.org/10.33387/jiko