Predictive Modeling of Air Quality Index Using Ensemble Learning and Multivariate Analysis

Anggina Primanita[a,1,*], Hadipurnawan Satria [b,2]

[a,b] Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Kec. Indralaya, Kabupaten Ogan Ilir, South Sumatera, 30862, Indonesia
[1] anggina@unsri.ac.id*; [2] hadi.ps@unsri.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 1 Oct 2024
Revised 12 Oct 2024
Accepted 17 Oct 2024

Keywords
Regression Algorithms
Air Quality Index
Random Forest
Decision Tree
K-Neural Network
Adaboost

ABSTRACT

Breathing polluted air can result in multiple health problems. Thus, it is important to understand and predict the air quality in the environment. Air Quality Index (AQI) is a unit used to measure the air pollutants. In Indonesia, this value is measured and published by the Meteorological, Climatological, and Geophysical Agency regularly. In this research, four commonly used regression algorithms were used to analyzed AQI data, namely, Random Forest, Decision Tree, K-Neural Network, and Ada Boost. All the algorithms model were developed to analyzed 1096 AQI data. The Mean Squared Error value of each model was computed as a measure of comparison. It is found that the Random Forest is the best performing algorithm. It can generalize well without overfitting to the data set.

1. Introduction

One of human's necessity is to breath. Breathing requires air that is readily available in the environment. However, in the later years, the air becoming polluted and it starts to threat human wellbeing [1]. It is found that breathing in a polluted air environment affecting human health in various factors, such as being the trigger for cardiovascular, respiratory diseases, and even related to lung cancer [2]. It is estimated that in 2022, polluted air is the main cause of 9 million deaths all around the world [3]. This number is likely to increase as the yearly review is reporting higher amount of pollutant in the air.

To measure the air pollutants, agencies and governments are using a standard called the Air Quality Index (AQI). It is a dimensionless unit that can inform people about the quality of air in a certain location. To measure the degree of pollutions in AQI, several measurements of pollutants are used, namely, the Particulate Matter (PM)2.5, PM10, Carbon Monoxide, Sulfur Dioxide, Nitrogen Dioxide, and Ozone, with most develop and some developing countries focusing on PM2.5 measurements [4].

In Indonesia, the measurement of AQI is focusing on PM2.5, as displayed at the official website of the Meteorological, Climatological, and Geophysical Agency (BMKG). However, there are several other data that is available as part of the air quality analysis, namely, PM2.5, PM10, Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), and Ozone (O3) [5]. These data are commonly measured at a fixed time interval, which results in a readily available public data.

Monitoring AQI becomes an important factor, especially in Indonesia, as there are more cities that is labeled as unhealthy by the Indonesian standard. The ability to predict the AQI will give more understanding of what to do to increase the air quality in Indonesia, which can affect decision making. Data analysis will also help decision maker to understand the main pollutant that affect Indonesia's AQI.

There are several algorithms that can be used to predict data. Regression algorithm is an umbrella term used to call algorithms that is modeled to analyze measured values on sample of observations [6] Some commonly used ones are Random Forest algorithm, Decision Tree algorithm, K-Neural Network algorithm, and Ada Boost. These algorithms are mainly used to understand the relationships between input and continuous output from a group of data. This relationship is then saved into a model that can be used to predict output from a new input that is previously unavailable [7]. Regression algorithms has been used to predict AQI in several countries, such as India[8], [9], [10] , China [11], Iran [12]. In its implementation, it is found that the algorithms have good performance.

Regression algorithms have also been used to predict AQI in Indonesia, however, we found that there are not many research that directly compares the quality of different models. Because of that, we are going to compare and analyze the performance of different regression algorithms to predict Indonesia's AQI.

2. Literature Study

a. Air Quality Index in Indonesia

Air Quality Index is an index that is used to measure the quality of breathable air in a certain location. The number can be based on several measurements namely the PM2.5, PM10, Carbon Monoxide, Sulfur Dioxide, Nitrogen Dioxide, and Ozone.

In Indonesia, the AQI measurement is mainly focused on PM2.5, however, the Meteorological, Climatological, and Geophysical Agency (BMKG) are also publishing another data as part of air quality analysis. These data are: PM10, Sulfur Dioxide, Nitrogen Dioxide, and Ozone. The AQI Index of a location in Indonesia can be measured by choosing the maximum value between all the measured data. The category of AQI in Indonesia is regulated by the government through the Ministerial Regulation No. P.14/MENLHK/SETJEN/KUM.1/7/2020. The standard can be seen on Table 1.

Table 1. Indonesia AQI Standard Category

Range	Category
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very Unhealthy
>300	Dangerous

b. Regression Algorithms

Regression algorithm is an umbrella term used to call algorithms that is modeled to analyze measured values on sample of observations [6] Some commonly used ones are Random Forest algorithm, Decision Tree algorithm, K-Neural Network algorithm, and Ada Boost. In this sub-chapter, we will describe each of the algorithms.

c. K-Nearest Neighbor Regressor

K-Nearest Neighbor (KNN) is a non-parametric machine learning algorithm that works by storing all the existing cases and classifies new cases by a certain distance function criterion [13]. This algorithm is used in many real-world cases because it does not give any assumption about the data distribution, thus, making it useful for cases with non-linear data. The KNN algorithm for regression is written as follows:

1. Choose the value of K: select the value of K, which is the number of nearest neighbors that will be considered.

2. Choose random K points: choose random K points from data points.
3. Calculate distances: calculate distance of each data points to selected K values.
4. Sort and select the K-Nearest Neigbor: sort the data points based on its distance to the K values and select the closest K.
5. Averaging: calculate average target value of the K-Nearest Neighbor.
6. Make a prediction: assign a predicted class or value to the new data point.
7. Evaluate the model: evaluate the model using evaluation methods to assess the performance.

d. Random Forest Regressor

Random forest algorithm is a supervised learning algorithm that is common to be used in regression problems. It is a term that is used to call ensemble methods that use tree-type classifiers[11]. In this algorithm, several models are independently making their predictions which will be aggregated in the end for its final prediction.

The Random Forest algorithm for regression is as follows [14]:

1. Define the number of decision trees (N)
2. For each of the tree, create a boostrap sample with different subset of data.
3. For each node in each tree, choose a random subset as candidate for splitting
4. For each splitting candidate, split the data on the best feature.
5. Grow each tree to maximum depth.
6. Aggregate the result of the trees by averaging the output
7. Evaluate the model.

e. Ada Boost Regressor

Adaptive Boosting or commonly known as Ada Boost is an ensemble learning algorithm. The boosting method used in Ada Boost is used to reduce bias and variance. It can convert multiple weak classifiers in order to create a strong one. This made the algorithm ideal for problem that requires high accuracy.

The Ada Boost algorithm for regression is as follows [10]:

1. Initialize weight of each training sample, the weight for each training sample is equal.
2. Train a weak classifier using a sample of the training data
3. Measure the weighted error the weak classifier
4. Determine the weight of the weak classifier
5. Update weight of each training sample
6. Repeat step 2-5 until error is minimized
7. Combine weak classifier into strong classifier.

f. Decision Tree

Decision Tree is a supervised learning algorithm that can be used for decision making. It uses tree data structure to represent data segmentation by applying a series of rules [12]. This algorithm gives a meaningful tree resulted from repetitive splitting. The resulting tree can be used for data exploration or finding the relation between input and output variables.

The Decision Tree algorithm can be written as follows:

1. Selecting the best attribute using a splitting metric such as Entropy.
2. Splitting the dataset based on (1).
3. Repeating the process on each subset to create a new node or leaf until the criteria is fulfilled.

g. Mean Squared Error (MSE)

Mean Squared Error (MSE) is an evaluation metric that can be used to measure the performance of a regression model. It calculates the square of the difference between the actual and predicted value. The formula of MSE is displayed in equation (1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - yp_i)^2 \quad (1)$$

With:

n = amount of testing data
y_i = actual output value
yp_i = predicted output value

3. Methodology

In this article, different methods of regression are used to create a prediction model for AQI in Indonesia. The framework of this research is displayed in **Fig. 1**.

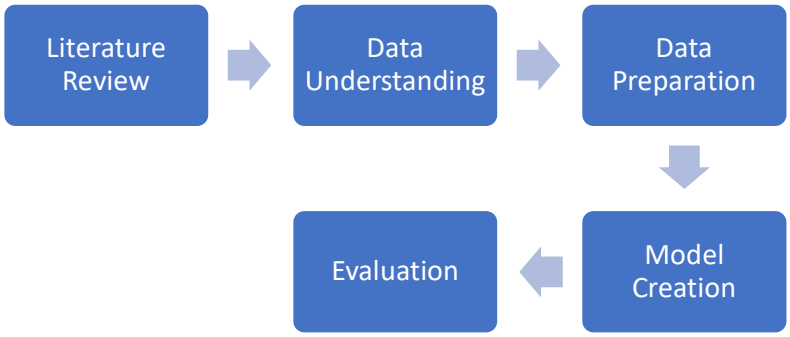

Fig. 1. Research Framework

Each of the steps were carried out to find the best model to predict AQI in Indonesia.

a. Data Understanding

In this step, the AQI data is obtained and analyzed. The data used in this research is from Kaggle. It is titled "Air Quality in South Tangerang, Indonesia 20-22". The data in the dataset is the daily measurement AQI in South Tangerang, Indonesia. The data consist of several measurement parameters that is displayed on **Table 2**.

Table 2. Dataset Information

Column Name	Type	Description
Date	Categorical	Date of the measurement
PM2.5	Numerical	Measurement result of Particulate Matter 2.5
PM10	Numerical	Measurement result of Particulate Matter 10
SO2	Numerical	Measurement result of Sulfur Dioxide
CO	Numerical	Measurement result of Carbon Monoxide
O3	Numerical	Measurement result of Ozone
NO2	Numerical	Measurement result of Nitrogen Dioxide
Max	Numerical	Highest value of the measurement during the day
Critical Component	Categorical	The measure with the highest value
Category	Categorical	AQI category based on the ministry regulation

To further understand the data, several processes was carried out. The first being finding data with null values. Based on the analysis it is found that there are 60 data with null values. As such, these data were dropped from the dataset.

The next process carried out was to visualize the categorical data using histogram. This is done to understand the distribution of AQI category present in the data. The histogram of the AQI categories is displayed in **Fig. 2**. In this stage, it can be seen that the data seems to be imbalanced, in which, the Good and Unhealthy category is significantly less than moderate.

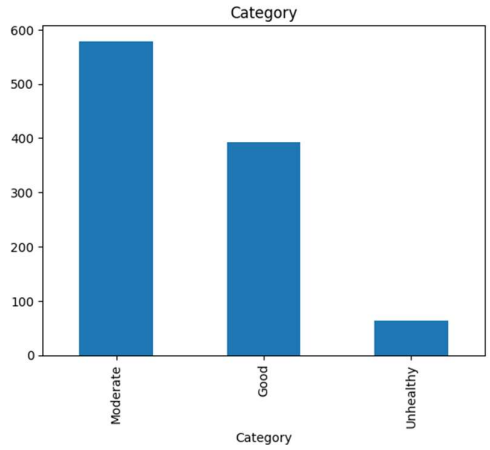

Fig. 2. Histogram of the data based on Category feature

The next analysis done in this phase is done by creating the correlation matrix to find whether all the columns correlate with the target output, in this case, the "Max" variable. The correlation matrix of this data is displayed on **Fig. 3**.

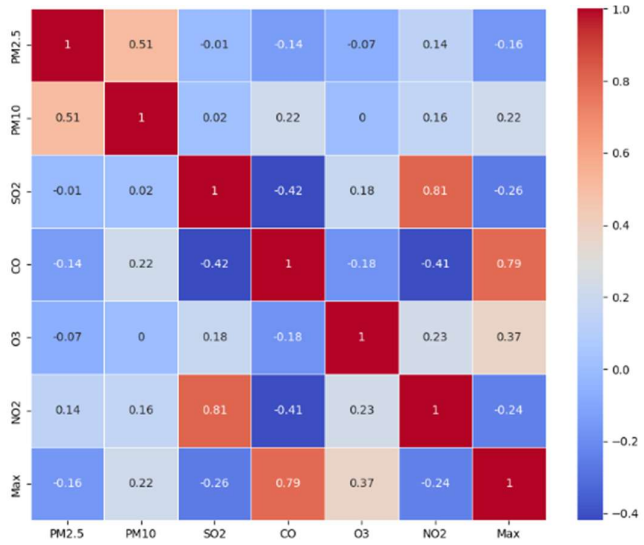

Fig. 3. Correlation matrix of the numerical features in AQI data

Based on the correlation matrix, it can be concluded that all the variables are correlated to the target variable to some degree, with the PM2.5, PM10, SO2 and NO2 correlating weakly; O3 correlating moderately; and CO correlating strongly to Max variable.

Upon examining the result of data analysis, some conclusion regarding the data was taken. The first is that the data is imbalanced, thus, it needs further processing to overcome it. Second, that all the measured variables are correlating to target variable, hence, all of the variables will be used.

b. Data Preparation

The next phase after understanding the data is to prepare it to fit the model that we will be using. In this phase, the categorical data will be encoded and an oversampling technique will be implemented in the encoded data to ensure data balance. In the end, the data will be split into train and test data.

There are three of what counts as categorical data from the dataset that will be encoded, namely, the Date, Critical Component, and Category features. The Date column from the dataset is still considered categorical because it combines day, month, and year. In this case, we are separating those data into 3 different columns. The Critical Component column is a column that contains categorical value of the maximum measurement of the measured day. This column in encoded using the One Hot Encoding method to ensure that the information is not lost.

The Synthetic Minority Oversampling Technique (SMOTE) is the oversampling technique that is used to increase overcome the imbalanced in the data set, mainly that the data from Unhealthy and Good category. This technique is known to be simple yet robust in adding new synthetic data to a dataset[15]. In this article, we implemented SMOTE using the "not_majority" parameter. Which means that the data will be resampled all classes except for the majority. Implementing SMOTE results in a more balanced data which can be seen in **Fig. 4**.

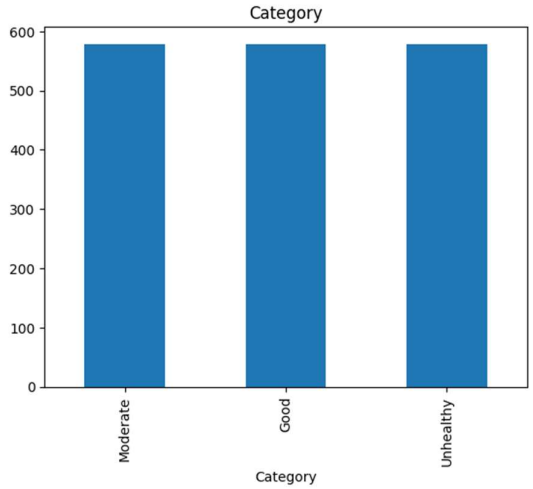

Fig. 4. Histogram of the data based on Category after SMOTE

After the oversampling process, the data is then split into the train and test data. The split was implemented in 80:20 ratio. In which 20% of the data (348) becomes the test dataset and the rest of it (1389 data) is part of the train data. The data from the train data is then normalized using the Standard Scaler. Following this, the data is deemed to be ready for training using the model that we will be using.

c. Model Creation

In this phase, a model from each regression algorithms that have been mentioned before will be created. The model is build using python and Sklearn libraries. The parameters were taken from studies in the literature study, as well as the recommended parameter from Sklearn documentation. The model and its parameters can be seen in Table 3.

Table 3. Model used to predict AQI and its parameters

Model	Parameter
K Neighbor Regressor	n_neighbor = 10
Random Forest Regressor	n_estimators = 50 max_depth = 16 random_state = 55
Ada Boost Regressor	learning rate = 0.05 random_state=55
Support Vector Regression	kernel = 'rbf' epsilon = 0.1

After all the models were build, each of the model is run independently using the train data that have been prepared. The resulting model is then used to calculate each of the models' MSE value using the train and test data. The MSE values is then stored to evaluate its result.

4. Result and Discussion

The result of each of the model is displayed in Table 4.

Table 4. MSE Values of each Regression Models

Model	MSE_Train	MSE_Test
K-Neighbor Regressor	0.035685	0.03726
Random Forest Regressor	0.000111	**0.000346**
Ada Boost Regressor	0.036502	0.036762
Decision Tree	**0.0**	0.000552

The K-Neighbor regressor have a similar MSE values on both training and test data. This shows that the model is able to generalize quite well. However, the MSE value of its test data is the highest amongst all of the model. This might comes from the lazy learner character of K-Neighbor Regressor. Which means that it might not be the best fit for data with complex pattern.

Random Forest Regressor is known to be a good model to predict data with many features. It shows in its MSE results, in which the value of MSE in train data is the second lowest after Decision Tree and the value of MSE in test data is the lowest. The base AQI data already has numerous numerical features with each of the measurements. Even more, there are added features from the One Hot Encoding process. After the data preparation phase, the AQI data has 21 features. It fits into the benefit of Random Forest Regressor.

The Ada Boost model also has similar MSE values, which shows that it is able to generalize the data without overfitting. The MSE value of its test data is the highest among the four models while the MSE value of its train data is the third highest. Which means that it has less, although comparable to K-Neighbor Regressor, ability to predict AQI data.

Decision Tree model has an MSE value 0.0 in its train data. This, although the lowest among the four model, indicates overfitting. Decision Tree is known for its ability to be able to handle many numerical and categorical data. However, based on this result. It is not a good fit for AQI data.

As seen in the MSE results, each of the model were considered as able to predict the AQI value, however, the results are varying. In MSE calculation, value closer to 0 is more desirable. In the table, it is clear that Decision Tree has the least MSE value of 0.0 when compared to the train data and Random Forest Regressor has the least MSE value of 0.000346 in the test data. The 0.0 MSE value that resulted from the Decision Tree indicates overfitting. Which means that the second best model with MSE value is the best fit to predict AQI data. This model is the Random Forest model.

5. Conclusion

In this research, four regression algorithms were used to analyze and predict the AQI data in South Tangerang, Indonesia. Before the algorithms were implemented, several processes were employed to make sure that the data that is feed into the algorithm is of good quality. The data has been cleaned of null values and had to be oversampled using SMOTE to overcome its imbalanced characteristic. Upon applying the regression models, it is found that the Random Forest is the best performing algorithm because of its low MSE. With 0.000111 for the training set and 0.000346 for the test set, it shows that it is able to generalize well without overfitting.

For future research, these models can be reused to analyze data in more cities in Indonesia. It would also be beneficial to expand the dataset to include more factors, such as weather conditions and traffic patterns to get a more comprehensive understanding of what causing the Air Quality Index and basis for prediction.

References

[1] L. Zhang and X. Ma, "A novel multi-fractional multivariate grey model for city air quality index prediction in China," *Expert Syst Appl*, vol. 257, Dec. 2024, doi: 10.1016/j.eswa.2024.125010.

[2] A. A. Almetwally, M. Bin-Jumah, and A. A. Allam, "Ambient air pollution and its influence on human health and welfare: an overview," *Environmental Science and Pollution Research*, vol. 27, no. 20, pp. 24815–24830, Jul. 2020, doi: 10.1007/s11356-020-09042-2.

[3] R. Fuller *et al.*, "Pollution and health: a progress update," Jun. 01, 2022, *Elsevier B.V.* doi: 10.1016/S2542-5196(22)00090-0.

[4] P. K and P. Kumar, "A critical evaluation of air quality index models (1960–2021)," *Environ Monit Assess*, vol. 194, no. 5, p. 324, May 2022, doi: 10.1007/s10661-022-09896-8.

[5] BMKG, "Official Website of Badan Meteorologi, Klimatologi dan Geofisika Indonesia," 2024. Last accessed: July 30, 2024.

[6] I. Pardoe, *Applied Regression Modeling*. Wiley, 2020. doi: 10.1002/9781119615941.

[7] F. Stulp and O. Sigaud, "Many regression algorithms, one unified model: A review," *Neural Networks*, vol. 69, pp. 60–79, Sep. 2015, doi: 10.1016/j.neunet.2015.05.005.

[8] S. Ketu, "Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India," *Natural Hazards*, vol. 114, no. 2, pp. 2109–2138, Nov. 2022, doi: 10.1007/s11069-022-05463-z.

[9] K. C. Atmakuri and K. V Prasad, "Urban Air Quality Analysis And Aqi Prediction Using Improved Knn Classifier," *Journal of Pharmaceutical Negative Results ¦*, vol. 13, 2022, doi: 10.47750/pnr.2022.13.S09.899.

[10] D. Thamizhselvi, B. Kasi, K. Kamalakkannan, S. Bharath, S. Gowtham, and M. A. Kishore, "Air Quality Prediction Using Adaboost," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICCEBS58601.2023.10448934.

[11] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ–A random forest approach for predicting air quality in urban sensing systems," *Sensors (Switzerland)*, vol. 16, no. 1, Jan. 2016, doi: 10.3390/s16010086.

[12] A. Jamal and R. N. Nodehi, "Predicting Air Quality Index Based On Meteorologi-Cal Data: A Comparison Of Regression Analysis, *Artificial Neural Networks and Decision Tree*," 2017. [Online]. Available: http://japh.tums.ac.ir

[13] B. BARAN, "Air Quality Index Prediction In Besiktas District By Artificial Neural Networks And K Nearest Neighbors," *Mühendislik Bilimleri ve Tasarım Dergisi*, vol. 9, no. 1, pp. 52–63, Mar. 2021, doi: 10.21923/jesd.671836.

[14] L. Breiman, "Random Forests," 2001.

[15] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," 2018.