Classification model of air quality in Jakarta using decision tree algorithm based on air pollutant standard index

F M Putra and I S Sitanggang

1Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University (Bogor Agricultural University), Bogor, Indonesia

*Corresponding author: fiqhri_mulianda@apps.ipb.ac.id

Abstract. The level of air quality is getting lower because of high levels of air pollution in big cities. Big cities in Indonesia also experience air pollution, this is caused by the increase in road users who use motorized vehicle materials, industrial development, land burning, waste accumulation so that air quality changes quite dramatically. Daily air quality needs to be accurately measured and classified. Accurate classification results will help the government in making policy. The aim is to control pollution to get air quality standards that can be useful for survival, especially in Jakarta. Air pollutants contain various components of elemental compounds such as carbon monoxide (CO), nitrogen dioxide (NO), sulfur dioxide (SO), particulate matter (PM), ozone (O) and nitrogen monoxide (NO). This study aims to determine the parameters that affect air quality in Jakarta using the C5.0 algorithm and Random Forest based on the Air Pollution Standard Index (ISPU) category. The classification algorithms used are C5.0 and Random Forest which are categorized in the Decision Tree model. C5.0 also produces rule-based models. The accuracy of the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2017 is 99.74%, 99.22%, and 99.97% with 1412 training data and 389 testing data. The accuracy of the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2018 is 98.28%, 98.85%, and 97.42% with 1439 training data and 349 testing data. The most important variable to classify air quality is Ozone (O) and air quality in Jakarta is dominated by the Moderate category in 2017 and 2018.

1. Introduction
Air quality issues have always been associated with the development of major cities in the world today. The level of air quality becomes lower every day due to high levels of air pollution in big cities. As well as major cities in Indonesia also experienced air pollution, this is due to increasing road users who use the vehicles, the construction industry, burning of land, accumulation of garbage so that the air quality changes quite dramatically [1]. Pollutants in the air contain various component elements compounds such as carbon monoxide (CO), nitrogen dioxide (NO), sulfur dioxide (SO), particulate matter (PM), ozone (O) and nitrogen monoxide (NO) [2]. When these compounds contain elements of high levels, it will affect people's health, especially in respiratory disorders and can also cause death.

Daily air quality needs to be accurately measured and classified. Accurate classification results will help the government in making policy. With the aim of controlling pollution so that it is at air quality standards that can benefit survival in big city such as Jakarta. The data mining algorithm has been applied to classify air pollution, one of which uses Artificial Neural Networks for the classification of
ambient air quality in Malaysia [3]. In 2015, Huang Min proposed a system by combining artificial neural networks and data mining for the prediction of air pollution data [4]. Furthermore, a study by Siregar reports air pollution levels using sensors in smart cities [5]. Another study applied the air quality classification in Pekanbaru using the K-NN algorithm with Euclidean distances based on categories on the air pollutant standard index (ISPU) [6]. In 2017 Ramadhan applied the classification was conducted to 8716 sample data from hotspot data in Sumatra 2014 and 9073 sample data from hotspot data in Sumatra 2015. Experimental results show that decision tree model of C5.0 provides the best accuracy in prediction compared to rule-based model of C5.0 and Random Forest model Accuracy of the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2014 is 96.8%, 96.0%, and 85.6%, respectively. Accuracy of the decision tree model and the rule-based model from C5.0 and Random Forest on the dataset of 2015 is 97.1%, 96.6%, and 75.6%, respectively [12].

The reason for using data mining is to find hidden information knowledge in databases that have large data with statistical calculations and mathematical models. Classification is a method to learn a target function $f$ that maps each attribute set $x$ to predefined class label $y$ in which the target function is also a known classification model [8]. The classification model used is a decision tree model and a rule-based model. A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [9]. Besides, a rule-based model is a set of if-then conditions that have been derived from a tree model into more simple conditions. Decision tree algorithms applied in this study are C5.0 and Random Forest, one of which with the best accuracy will be selected for creating the best classification model.

C5.0 is a method in decision tree works by testing the classifier first to classify unseen data and for this purpose resulting decision is used. Pandya and Pandya in their study conclusively explain that C5.0 is an improvement of the C4.5 algorithm which is faster in processing time, more efficient in memory usage, lower in error, and ultimately more accurate for classification [10]. The algorithm Random Forest (RF) is a method for generating a child node to every node randomized thereby increasing the accuracy of the results. This Random Forest is used to create a decision tree consisting of a root node, internal nodes, and leaf nodes by taking random attribute data and by prevailing regulations. The root node is the node that is located at the top or commonly referred to as the root of the decision tree. An internal node is a branch node, in which this node has only one input and output at least two and. While the leaf node or a terminal node is the last node that has only one input and has no output. The decision tree begins by calculating the entropy values as a determinant of impurity levels of attributes and value of information gain [10].

Index of air quality standards is used officially in Indonesia is by the Decree of the Minister of the Environment Number: KEP 45 / MENLH / 1997 on Air Pollutant Standard Index (ISPU). In the decision which is used as information in the form of numbers that do not have a unit describing the condition of the air quality to the public on ambient air quality in the given time and location [2]. Parameters of air pollution includes carbon monoxide (CO), nitrogen dioxide (NO$_2$), sulfur dioxide (SO$_2$), particulate matter (PM$_{10}$), ozone (O$_3$) and nitrogen monoxide (NO) from five air stations in Jakarta [2]. This study applied a data mining technique namely classification using the C5.0 and Random Forest algorithms on air pollutant standard index dataset in Jakarta.

2. Methods

2.1. Data Collection

The dataset used in this study is Jakarta's water quality in 2017 and 2018 collected from the Jakarta Environment Agency with a frequency of 1 month publication measurements from 7am to 8am in CSV format. This dataset contains the Air Pollutant Standard Index measured from five air quality monitoring stations consisting of 10 attributes: date, stations, CO, PM$_{10}$, NO$_2$, O$_3$, SO$_2$, max, critical and category [11].
2.2. Data Pre-processing
After the data were collected, then the pre-processing stage was carried out including data selection and data cleansing. In the data selection step, from 10 existing attributes, six related attributes were selected, namely CO, PM$_1$, NO$_2$, O$_3$, SO$_2$, and the category. Number of objects is 1826 records of 2017 and 1,826 records of 2018. At the data cleaning step, missing values were replaced by averaging all value of an attribute. Number of objects after cleaning step is 1801 records of 2017 and 1788 records of 2018.

2.3. Model Development
At this stage, the model was created using the C5.0 algorithm and the Random Forest algorithm. Furthermore, performance of the model was measured to find out how accurately the model to classify this Air Pollutant Standard Index. This step was carried out using R programming language. The classification using the C5.0 algorithm was implemented using the library C5.0 and the Random Forest algorithm was implemented using the library random forest.

2.4. Testing Model
Calculation of accuracy plays an important role so that the results obtained are considered as success and failure. The accuracy of classifier can be measured based on a confusion matrix by analysing how well the classifier recognizes tuples from different classes. TP and TN provide information when the classifier is correct, while FP and FN provide information when the classifier is incorrect [9]. Figure 1 is an example of a confusion matrix.

| Predictive class | Actual class |
|------------------|-------------|
|                  | Yes | No |
| Yes              | TP  | FN |
| No               | FP  | TN |
| Total            | P'  | N' |

**Figure 1.** Confusion Matrix (CM)

Accuracy is the percentage of correctly predicted data. The calculation of accuracy in Equation 1 is as follows:

$$A = \frac{(T + T)}{(T + T + F + F)}$$

where:
- TP: True positive, is the number of data with positive class classified as positive.
- TN: True negative, is the number of data with negative class classified as negative.
- FP: False positive, is the number of data with positive class classified as negative.
- FN: False negative, is the number of data with negative class classified as positive.

3. Results and Discussion
The training and testing data distribution was done to see the accuracy of the classification results. The dataset was divided into 80% for training data and 20% for testing data. Number of sample data of air quality dataset 2017 are 1801 with 1412 training data and 389 testing data. Number of sample data of air quality dataset 2018 are 1788 with 1439 training data and 349 testing data.

The C5.0 algorithm results two classification model namely decision tree model and rule-based model. The decision tree in Figure 2 shows that the root node is $G_2$, the internal position of the first node is “$G_2 > 100$” with the Unhealthy category, while the second node has label “$G_3 < 197$” with a Moderate category. The decision tree in Figure 3 shows that the root node is $G_3$, the internal position of the first
node is “$G_3>197$” with the Very Unhealthy category, while the second node has the label “$G_3<197$” with an Unhealthy category. Based on the classification results, all parameters affect air quality in Jakarta. Table 2 provides the confusion matrix of decision tree model 2017 whereas Table 3 provides the confusion matrix of decision tree model 2018. Based on the confusion matrix in Table 2 and Table 3, the accuracy of the model 2017 reach 99.74% whereas the accuracy of the model 2018 is 98.28%.

Figure 2. Decision tree of Jakarta air quality dataset 2017

Figure 3. Decision tree of Jakarta air quality dataset 2018
Table 1. Confusion matrix of decision tree model 2017

| Actual class | Good | Moderate | Unhealthy |
|--------------|------|----------|-----------|
| Data Training Predictive class | Good | 368 | 0 | 0 |
|                     Moderate | 0 | 888 | 0 |
|                     Unhealthy | 0 | 0 | 156 |
| Data Testing Predictive class | Good | 113 | 0 | 0 |
|                     Moderate | 1 | 233 | 0 |
|                     Unhealthy | 0 | 0 | 42 |

Table 3. Confusion matrix of decision tree model 2018

| Actual class | Good | Very | Moderate | Unhealthy |
|--------------|------|------|----------|-----------|
| Data Training Predictive class | Good | 298 | 0 | 2 | 0 |
|                     Very Unhealthy | 0 | 29 | 0 | 0 |
|                     Unhealthy | 0 | 0 | 300 | 0 |
| Data Testing Predictive class | Good | 66 | 0 | 0 | 0 |
|                     Very Unhealthy | 0 | 1 | 0 | 0 |
|                     Moderate | 4 | 0 | 197 | 0 |
|                     Unhealthy | 0 | 0 | 0 | 79 |

Rule-based model of Air Quality Index 2017 and 2018 is given in Table 4. The rule-based model in Table 4 shows all parameters affecting air quality in Jakarta. From the above rules, it was found that the Air Pollutant Standard Index in Jakarta in 2017 and 2018 was in the Moderate category. A rule-based model was tested using the training data and testing data. The result is presented in confusion matrix.

Table 4. Rule-based model of Air Quality Index 2017 and 2018

| Rule 1: (275/1, elevator 4.8) | Rule 1: (356, lift 3.8) |
|-------------------------------|------------------------|
| PM10 <= 50                    | pm10 <= 50             |
| SO2 <= 49                     | so2 <= 50              |
| o3 <= 50                      | co <= 49               |
| -> Class GOOD [0993]          | o3 <= 50               |
|                               | -> class GOOD [0.997]  |

Rule 2: (10 lifts 4.4)
| Rule 2: (128, lift 1.6) |
|-------------------------|
| PM10> 52                | pm10 > 50              |
| PM10 <= 52.36606        | o3 <= 50               |
| o3 <= 50                | -> class MODERATE [0.992] |
| -> Class GOOD [0917]    |                       |

Rule 3: (15/1, elevator 4.3)
| Rule 3: (749/12, lift 1.6) |
|-----------------------------|
| PM10 <= 50                  | pm10 <= 98             |
| co <= 47                    | o3 > 50                |
| o3> 76                      | o3 <= 100              |
| -> class MODERATE [0.983]   | -> class MODERATE [0.983] |

Rule 4: (35, lift 1.5)
Table 5 provides the confusion matrix for rule-based model of 2017. The confusion matrix for rule-based model of 2018 is presented in Table 6. Based on the confusion matrix in Table 5 and Table 6, the accuracy of the 2017 model reaches 99.22% while the accuracy of the 2018 model is 98.85%.

**Table 5. Confusion matrix for rule-based model of 2017**

| Actual class   | Data Training | Predictive class | Good | Moderate | Unhealthy |
|----------------|---------------|------------------|------|----------|-----------|
|                |               |                  | 356  | 0        | 0         |
| Good           | 12            | 888              | 0    | 0        | 156       |
| Moderate       | 0             | 0                | 0    | 42       | 0         |

**Table 6. Confusion matrix for rule-based model of 2018**

| Actual class   | Data Testing | Predictive class | Good | Moderate | Unhealthy |
|----------------|--------------|------------------|------|----------|-----------|
|                | 111           | 0                | 0    | 0        | 42        |
| Good           | 3             | 233              | 0    | 0        | 0         |
| Moderate       | 0             | 0                | 0    | 42       | 0         |
| Unhealthy      | 0             | 0                | 0    | 0        | 42        |
Table 6. Confusion matrix for rule-based model of 2018

| Actual class | Good | Very Unhealthy | Moderate | Unhealthy |
|--------------|------|----------------|----------|-----------|
| Data Training Predictive class Good | 298  | 0              | 1        | 0         |
| Very Unhealthy | 0    | 29             | 0        | 0         |
| Moderate      | 0    | 0              | 811      | 0         |
| Unhealthy     | 0    | 0              | 0        | 300       |
| Data Testing  Predictive class Good | 66   | 0              | 0        | 0         |
| Very Unhealthy | 0    | 1              | 0        | 0         |
| Moderate      | 4    | 0              | 199      | 0         |
| Unhealthy     | 0    | 0              | 0        | 79        |

This study applied the Random forest algorithm in addition to the C5.0 algorithm. Table 7 and Table 8 provide confusion matrix of rule-based models on the dataset of 2017 and 2018 respectively. Based on the confusion matrix in Table 7 and Table 8, the accuracy of the model on the dataset 2017 reaches 99.97% while the accuracy of the model on the dataset 2018 is 97.42%.

Table 7. Confusion matrix of random forest model on the dataset 2017

| Actual class | Good | Moderate | Unhealthy |
|--------------|------|----------|-----------|
| Data Training Predictive class Good | 366  | 0        | 0         |
| Moderate      | 2    | 888      | 0         |
| Unhealthy     | 0    | 0        | 156       |
| Data Testing  Predictive class Good | 110  | 0        | 0         |
| Moderate      | 4    | 233      | 0         |
| Unhealthy     | 0    | 0        | 42        |

Table 8. Confusion matrix of random forest model on the dataset 2018

| Actual class | Good | Very Unhealthy | Moderate | Unhealthy |
|--------------|------|----------------|----------|-----------|
| Data Training Predictive class Good | 295  | 0              | 1        | 0         |
| Very Unhealthy | 0    | 29             | 0        | 0         |
| Moderate      | 3    | 0              | 812      | 0         |
| Unhealthy     | 0    | 0              | 0        | 300       |
| Data Testing  Predictive class Good | 61   | 0              | 0        | 0         |
| Very Unhealthy | 0    | 1              | 0        | 0         |
| Moderate      | 9    | 0              | 199      | 0         |
| Unhealthy     | 0    | 0              | 0        | 79        |

Training process of Air Quality Index dataset of 2017 and 2018 is given in Figure 4 and Figure 5 respectively. In order to get more optimal results, Random forest creates 300 trees on the Air Quality
Index dataset of 2017 and 2018. In Figure 4 and Figure 5 it can be seen that the plot produced in the tree model of up to 200 training has been stable so that the model can be used for data classification.

**Figure 4.** The training process of random forest on the dataset 2017

**Figure 5.** The training process of random forest on the dataset 2018
Figure 6 shows that the important variables that highly influence the classification of the Air Pollutant Index in Jakarta are Ozone ($O_3$) which occupies the root node on the tree followed by PM$_{10}$, SO$_2$, NO$_2$, and CO.

This study implemented two classification algorithms namely C5.0 and Random Forest. The accuracy of decision tree model and rule-based model as the results of C5.0 algorithm as well as the accuracy of model generated from the Random Forest are provided in Table 9. The results show that both C5.0 and Random Forest are working properly in classifying Air Quality Index dataset of 2017 and 2018 in which the accuracy of classification models is greater than 90%.

| Model              | C5.0     | Random forest |
|--------------------|----------|---------------|
|                    | decision tree | rule-based | Dataset 2017 | Dataset 2018 | Dataset 2017 | Dataset 2018 |
| Training set       | 100% | 98.86% | 99.15% | 99.93% | 99.85% | 99.79% |
| Testing set        | 99.74% | 98.28% | 99.22% | 98.85% | 99.97% | 97.42% |
4. Conclusion

This study implemented two classification algorithms in data mining namely C5.0 and Random Forest in order to generate a model for air quality prediction on Air Pollutant Standard Index datasets in Jakarta in 2017 and 2018. Experimental results show that the Random Forest model provides the best accuracy in prediction on dataset 2017 compared to the rule-based model of C5.0 and the decision tree model of C5.0. The rule-based model of C5.0 provides the best accuracy in prediction on dataset 2018 compared to the decision tree model of C5.0 and Random Forest model. The important variables that influences the classification of the Air Pollutant Index in Jakarta are Ozone ($O_3$), P, S, N, and CO. In addition, the result shows that air quality in Jakarta is dominated by the Moderate category in 2017 and 2018.

References

[1] Dede N S Hanif F Napitupulu L Soejachmoen M H Murhajanti P Syahril S M F and Husin S 2003 *Udara Bersih Hak Kita Bersama* (in Bahasa) (Jakarta: Pelangi)
[2] DLHK 2013 Laporan Tahunan Kualitas Udara sekitar Kota Pekanbaru (in Bahasa) Laboratorium Udara
[3] Kamal M M Gilani R Shauri R L A 2006 Prediction of ambient air quality based on neural network technique 4th *Student Conference on Research and Development (SCORED)* pp 115-119
[4] Huang M Zhang Q Wang J Zhu L 2015 A new air quality forecasting model using data mining and artificial neural network 6th *IEEE International Conference on In Software Engineering and Service Science (ICSESS)* pp 259-262
[5] Siregar B Nasution A B A Fahmi F 2016 Integrated pollution monitoring system for a smart city *Int. Conf. on ICT For Smart Society (ICISS)* pp 49-52
[6] Arief A 2014 Klasifikasi Kualitas Udara Pekanbaru Menggunakan Algoritma K-NN Dengan Euclidean Distance Berdasarkan Kategori Indeks Standar Pencemaran Udara (ISPU) *Skripsi Teknik Informatika UIN Suska Riau* (in Bahasa)
[7] Badan Pengendalian Dampak Lingkungan Pedoman Teknis Perhitungan dan Pelaporan Serta Informasi Indeks Standar Pencemar Udara http://www.cetsuui.org/BML/Udara/ISPU/ISPU%20(Indeks%20Standar%20Pencemar%20Udara).htm (Accessed on April 12 2019) (in Bahasa)
[8] Tan P N Steinbach M Kumar V 2006 *Introduction to Data Mining* (Boston: Pearson Addison Wesley)
[9] Han J Kamber M Pei J 2012 *Data Mining Concepts and Techniques Third Edition* (USA: Morgan Kaufmann Publisher)
[10] Pandya R and Pandya J 2015 C5.0 algorithm to improved decision tree with feature selection and reduced error pruning *Int. Journal of Computer Applications* 177 16
[11] Dinas Lingkungan Hidup Provinsi DKI Jakarta 2018 Indeks Pencemar Udara (ISPU) di Provinsi DKI Jakarta Tahun 2018 http://data.jakarta.go.id/dataset/indeks-standar-pencemar-udara-di-provinsi-dki-jakarta-tahun-2018[Accessed: April 10 2019] (in Bahasa)
[12] Ramadhin M M, Sitanggang I S and Anzani L P 2017 Classification model for hotspot sequences as indicator for peatland fires using data mining approach *Matter: International Journal of Science and Technology* Special Issue 3(2) 588-597