Predicting Air Pollution Levels Using the Decision Tree Algorithm

Dosen Pembimbing:

1. Dr. ADITYA KURNIAWAN, S.Kom., MMSI., CND, CEHmaster
2. Chrisando Ryan Pardomuan Siahaan S.Kom, M.Kom.

1st Jacky Tanio

Binus University

Tangerang, Indonesia

[jacky.tanio@binus.ac.id](mailto:jacky.tanio@binus.ac.id)

BINUS UNIVERSITY  
2024

2nd Steven

Binus University

Tangerang, Indonesia  
[steven105@binus.ac.id](mailto:steven105@binus.ac.id)

***Abstract*— In the contemporary era, increasing levels of air pollution have become an urgent global concern, significantly affecting environmental sustainability and public health. Taking on this challenge requires an innovative approach based on advanced technologies and analytical methodologies. With the help of Machine Learning Algorithms like Decision Trees, we can provide systems to improve air quality. Decision Trees, one of the algorithms renowned for their interpretability and predictive capabilities, offer a promising path to develop robust strategies for reducing air pollution. Through comprehensive analysis and experiments, this research aims to reveal the potential of the Decision Tree algorithm in reducing air pollution, thereby encouraging a healthier and more sustainable environment.**

***Keywords— Decision Tree, Air Pollution, Algorithm, Machine Learning***

1. Introduction (*Heading 1*)

Air is an element of mixed or collection of gases on the surface of earth. Air plays a very important role because air can maintain all of the life creatures or living things on earth. Without air there would be no living things because our metabolic processes that occur in our bodies cannot take place without oxygen from the air because metabolic processes need air to do its process .

According to WHO, air pollution causes around 7 million deaths every year and contributes to various health problems such as respiratory diseases, cardiovascular diseases and cancer. Air pollution affects the air quality in a place, making the area degraded and less good for living things, and can even become dangerous. Many factors trigger a decrease in air quality in an area, including transportation, industrial, service activities, etc. which increase residual activity emissions into the air (Sani, 2016). In big cities like DKI Jakarta, air pollution is a serious problem faced by the people. Air pollution problems in big cities like DKI Jakarta are very serious because air pollution could cause health problems, especially respiratory problems like asthma, Chronic Obstructive Pulmonary Disease and long-term health problems like cardiovascular disease, lung cancer, and many more respiratory health problems. If this continues it could cause a lot of problems. Air pollution can also affect environmental quality, reduce visibility, and damage ecosystems (Rahmawati, 2016) [1].

Classification is the process of building models or classification rules using training data as a learning stage and testing data as a testing stage so that classification is included in the supervised learning method. Model construction is based on the characteristics of each existing object and groups the objects into appropriate classes (Wibawa et al., 2018).

Decision Trees are a method that falls under classification. A Decision Tree is a flow diagram shaped like a tree and each branch of the tree states a result. This method has tree nodes that represent the variables that have been tested and the branches that are formed are the division of test results and leaf nodes that represent certain class groups (Setio et al., 2020). Decision Trees are built top-down (from top to bottom) where the initial stage is carried out by evaluating all existing variables using statistical measures to measure the effectiveness of variables in classifying sample data sets (Kasih, 2019) [2]. The decision tree algorithm applied in this study are C5.0 and Random Forest, because C5.0 and Random Forest are the one with the best accuracy which will be chosen to create the best classification model [3].

C5.0 is one of the methods of decision tree work by testing classifier data to classify hidden data that are hidden from the classifier data and this purpose resulting in the decision to be used. Pandya and Pandya explain that from their study C5.0 is an improvement from the C4.5 algorithm which C5.0 is better in processing time because its methods that make data processing time faster than C4.5 method, then C.5 method use less memory usage which is a improvement from C4.5 because it use lesser memory usage, and the last one is more accurate for classification which is better to use this method than the C4.5. Random Forest (RF) algorithm is a method to generate child nodes from decision trees to every randomized node which by using this method could increase the accuracy of the result. This Random Forest is used to create a decision tree algorithm consisting of a root node, internal nodes, and leaf nodes by taking random data and by prevailing regulations. The root node is located at the top node or commonly referred to as the root of the decision tree. An internal node is a branch node in which this node only has one input and only one output. While the leaf node or a terminal node is the last node that only has one input but doesn't have output. Usually decision tree begins by calculating all of the entropy values first as a determinant of impurity levels of attributes and value of information gain [3]

In Indonesia, Index of air quality standards are used by the Decree of the Minister of the Environment on the Air Pollutant Standard Index (ISPU). Decision tree algorithms are used as information in the form of numbers that do not have a unit to describe the condition of the air quality to the public on ambient air quality in the given time and location [2]. There are parameters of air pollution which include carbon monoxide (CO), nitrogen dioxide (NO ), sulfur dioxide (SO ), particulate matter 2.5 (PM 2.5), particulate matter 10 (PM 10) , ozone (O ) and nitrogen monoxide (NO) from five air stations in Jakarta [2]. This study is applied by using data mining techniques, namely all classification data using the C5.0 and Random Forest algorithms on air pollutant standard index dataset in Jakarta [3].

Our goal in writing this article is to develop air pollution prevention to the point where pollution can be eliminated almost completely. By relying on the Decision Tree algorithm and direct measurement of pollution levels. We hope to reduce air pollutants significantly so that fewer people suffer from respiratory diseases and environmental damage.

# Previous Research

In this research, we discuss reducing air pollution using decision tree algorithm analysis. The main reason for this research is because air pollution is currently getting worse and can have an impact on human health. This research is not only carried out by researchers, but also by government authorities and industry around the world because air pollution is a global problem that is very dangerous because it kills living things slowly. Therefore, the main problem of this research is air pollution.

In this study, they used a data flow diagram for the proposed air pollution reduction approach. The data consists of three process steps: data extraction, transformation & loading, analytics, and knowledge discovery by using a rule base method. The first step of the research is data extraction - transformation & loading where they fetch the required data from the outside world through various data sources and use Pollution Master DB to store all the diverse data.

The second step of this research is analytics using Decision Trees to predict pollution levels in the coming days. The tools they use are Pollution Master DB and Historical DB because both are input to produce Decision Trees. Using these two tools, sample data can be generated based on the input. The rule base is also derived from data analysis which serves as a solution to reduce pollution. The reason they use it is because Knowledge DB can store information about upcoming events that might affect pollution levels in the city.

The final step of this research is knowledge discovery using a rule base where citizens can visualize predictions for the coming days and they can reduce air pollution by carrying out all the solutions that have been suggested by the analytics and also need the help of the government to convey information or create rules to help reducing air pollution..

Beside using the Decision Tree Algorithm in this research they also use A Supervised Machine-Learning Approach which helps them to create a predictive models by observing a system with minimum human intervention, so it is a really helpful tool to get all of the air pollution data.[8] The reason that in this research they also use machine learning is because of its effectiveness for getting all of the inputs and continuous measurement for an observation.[4]

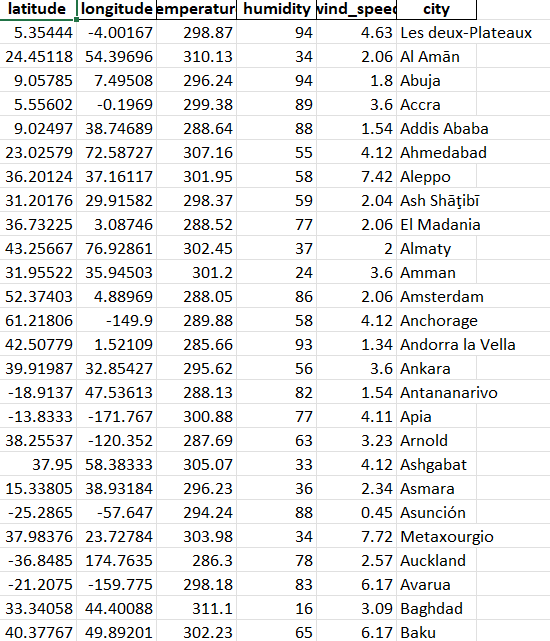
In this research they use Decision Tree Algorithm because Decision trees have several advantages that are better for predicting data such as interpretability compared with other methods. They are appropriate for handling both categorical and quantitative. The researchers also use ID3 (Iterative Dichotomiser 3) Algorithm to help the air pollution measurement and by using this algorithm they could measure all of the data set.

The research mainly focused on measuring the air pollution by using different kinds of methods and all of the methods that they use, they implemented in the decision tree to make better measurements and also to find the solutions. [2]

1. Research Methodology

This research is based on quantitative research because the results are based on quantitative results and also use data to find research results. The data source used in this research is the OpenWeather API, where information is collected in JSON format. Global weather and air quality data provided by the OpenWeather API can be accessed using specific geographic coordinates, such as the latitude and longitude of the city being analyzed. Then, this data can be used to project air pollution levels based on temperature, humidity and wind speed.. Using OpenWeatherAPI, we can predict air pollution in the city. Because apart from the effects of air pollution on our world, it also affects our health, especially our lungs because every time we breathe air, there are pollutants that enter our body, especially the lungs [7].

In this research to find the air pollution we must use geographic location such as the latitude and the longitude of the city coordinate to specify the prediction of the air pollution, the example is like we use Jakarta coordination Longitude: 106.8272, Latitude: -6.175 beside to find the air pollution data we use three kinds of parameter which are Temperature, Humidity, and Wind Speed to predict the current air pollution data in the specific city that we want to know by using the geographic coordinate[6]. The overview of our data set is like this.



In this research we use The decision tree algorithm to help predict air pollution data because by using The decision tree algorithm The prediction of the specific city that we want to predict can be more accurate rather than not using The Decision Tree algorithm.[5].

1. Result and Evaluation

In this paper, the results will be based on three parameters: Temperature, Wind Speed, and Humidity. These three parameters will be used to predict pollution levels in certain cities in a country. The dataset we use is taken from OpenWeatherAPI <https://openweathermap.org/api/air-pollution>.

As mentioned, from the 200.000 data set we will use only three different cities namely Jakarta, Tokyo, and New York as samples. Below is a sample table that we use to predict air pollution levels. The table consists of temperature, wind speed, and humidity for each city [9].

Table 1

| Parameter | Jakarta | Tokyo | New York |
| --- | --- | --- | --- |
| *Longitude* | 106.8272 | 139.6917 | -74.006 |
| *Latitude* | -6.1754 | 35.6895 | 40.7143 |
| *Humidity* | 83 | 87 | 32 |
| *Wind Speed* | 2.06 | 4.63 | 6.69 |
| *Temperature* | 27,38°C | 15,42°C | 13,71°C |

Here are the result of Air Pollution from Jakarta, Tokyo, and New York.

Based on the results and analysis, it can be said that wind speed influences air pollution levels more than temperature and humidity. Even with higher humidity, wind speed in Tokyo can affect air pollution more compared to Jakarta.

Based on the experimental results above, it can be concluded that humidity, wind speed and temperature influence air pollution. The higher the wind speed and the lower the temperature and humidity a city has, the more likely it is that air pollution will have higher predictions compared to a city with high temperature and high humidity with low wind speed. This is because higher wind speeds can spread pollutants over a wider area instead of allowing them to settle or dilute them. Meanwhile, low temperatures can cause temperature inversions where layers of warm air trap cold air and pollutants near the ground surface, which can cause the buildup of pollutants near the surface and worsen pollution levels. Lastly, low humidity can keep particles suspended in the air for longer periods. This can increase the total concentration of pollutants in the air. Thus, this paper shows the relationship between humidity, wind speed, and temperature that can influence air pollution levels. This paper serves as proof that the above statement can achieve success.[10].

1. Conclusion

The conclusion of this research is that the Decision Tree algorithm can be used effectively to predict air pollution levels using three main parameters: temperature, air speed, and humidity level. Although these parameters are not directly related to air pollution, this research shows that they can provide significant clues to pollution levels. For example, increasing temperatures can increase chemical reactions that produce ozone, while higher wind speeds can help disperse pollutants more widely and reduce their concentrations in certain areas.

Apart from that, this research also reveals that low humidity can increase the time pollutant particles remain in the air. This causes the total concentration of pollutants to be higher. As a result, it can be predicted well the complex relationship between temperature, humidity, wind speed, and air pollution levels using the Decision Tree algorithm.

Besides, one of the advantages is the use of the OpenWeather API to collect real-time and historical weather and air pollution data. This is very important in making accurate predictions. Using JSON format to retrieve data from the OpenWeather API makes further processing and analysis easier.

Findings from the research show that the influence of wind speed on air pollution levels is more significant than temperature and humidity. Even though Jakarta has lower humidity, air pollution levels can be significantly affected by wind speed in Tokyo.

Based on experiments, it can be concluded that predictions of air pollution levels can be better by combining data on humidity, wind speed and temperature. Air pollution levels in a city tend to increase when wind speeds are high and temperature and humidity are low. This is caused by high wind speeds which can spread pollutants widely, as well as low temperatures which can create temperature inversions where layers of warm air are trapped with cold air and pollutants near the ground surface.

This research proves that the Decision Tree algorithm has great potential in predicting air pollution levels and can provide valuable information to the government and society to take appropriate preventive measures. However, additional research is needed to explore other methods and factors that may influence air pollution predictions more accurately.

##### REFERENCES

1. Wiranata, A. D., Soleman, S., Irwansyah, I., Sudaryana, I. K., & Rizal, R. (2023, June 30). KLASIFIKASI DATA MINING UNTUK MENENTUKAN KUALITAS UDARA DI PROVINSI DKI JAKARTA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS (K-NN). *Infotech: Journal of Technology Information*, *9*(1), 95–100. <https://doi.org/10.37365/jti.v9i1.164>
2. Fahira, A.A. 2023. PERBANDINGAN METODE K-NEAREST NEIGHBOR DAN DECISION TREE UNTUK KLASIFIKASI KUALITAS UDARA BERDASARKAN BOOTSTRAP SAMPLING. Thesis. Universitas Sriwijaya. Palembang
3. Putra, F. M., & Sitanggang, I. S. (2020, July 1). Classification model of air quality in Jakarta using decision tree algorithm based on air pollutant standard index. *IOP Conference Series: Earth and Environmental Science*, *528*(1), 012053. <https://doi.org/10.1088/1755-1315/528/1/012053>
4. Madan, T., Sagar, S., & Virmani, D. (2020, December 18). Air Quality Prediction using Machine Learning Algorithms –A Review. *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. https://doi.org/10.1109/icacccn51052.2020.9362912
5. Althuwaynee, O. F., Balogun, A. L., Aydda, A., & Gumbo, T. (2020, January 1). Classification of air pollutants API Inter-Correlation using decision tree algorithms. *IOP Conference Series: Earth and Environmental Science*, *419*(1), 012022. <https://doi.org/10.1088/1755-1315/419/1/012022>
6. Kim, B. Y., Lim, Y. K., & Cha, J. W. (2022, October). Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms. *Atmospheric Pollution Research*, *13*(10), 101547. <https://doi.org/10.1016/j.apr.2022.101547>
7. Althuwaynee, O. F., Balogun, A., & Al Madhoun, W. (2020, January 10). Air pollution hazard assessment using decision tree algorithms and bivariate probability cluster polar function: evaluating inter-correlation clusters of PM10 and other air pollutants. *GIScience & Remote Sensing*, *57*(2), 207–226. <https://doi.org/10.1080/15481603.2020.1712064>
8. Halsana, S. (2020, July 20). Air Quality Prediction Model using Supervised Machine Learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 190–201. <https://doi.org/10.32628/cseit206435>
9. Hussain, M., Afrin, S., Irin, A., & Park, S. K. (2021, December 17). Applying Decision Tree Algorithm for Air Quality Prediction in Bangladesh. *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*. https://doi.org/10.1109/eict54103.2021.9733443
10. Alaba, O., Ogunsanwo, G., Olayiwola, O., Taiwo, E., & Abass, O. (2021). Effective Analysis of Air Pollution Using Decision Tree, Naive Bayes and Zeror Classifiers. *Journal of Science and Information Technology*, *16*(1), 37–50. Retrieved from http://journals.tasued.edu.ng/index.php/josit/article/view/31