JNEST EISSN: 2961-8738 ● PISSN: 2961-8916 ● DOI: 10.56741/jnest.v2i01.314
Article Received 20 March 2023. Revised 02 April 2023.

AI Big Data System to Predict Air Quality for Environmental Toxicology Monitoring

Adi Jufriansah*, Azmi Khusnani, Yudhiakto Pramudya, Nursina Sya'bania, Kristina Theresia Leto, Hamzarudin Hikmatiar, Sabarudin Saputra

***Abstract*—Pollutants in the air have a detrimental effect on both human existence and the environment. Because it is closely linked to climate change and the effects of global warming, research on air quality is currently receiving attention from a variety of disciplines. The science of forecasting air quality has evolved over time, and the actions of different gases (hazardous elements) and other components directly affect the health of the ecosystem. This study aims to present the development of a prediction system based on artificial intelligence models using a database of air quality sensors.This study develops a prediction model using machine learning (ML) and a Decision Tree (DT) algorithm that can enable decision harmonization across different industries with high accuracy. Based on pollutant levels and the classification outcomes from each cluster's analysis, statistical forecasting findings with a model accuracy of 0.95 have been achieved. This may act as a guiding factor in the development of air quality policies that address global consequences, international rescue efforts, and the preservation of the gap in air quality index standardization.**

***Index Terms*—Air quality, Artificial intelligence, Environmental toxicology, Prediction systems**

I. INTRODUCTION

THEscientific field of toxicology investigates how chemicals affect both the environment and living objects. Air pollutants that are challenging to regulate may be connected to toxicology. Environmental toxicology is typically brought on by a decline in environmental quality, which is brought on by an increase in toxic chemicals along with the advancement of industrial technology. The disruption of organismal processes is impacted by this factor. According to the Air Quality Live Index (AQLI), the City of Jakarta has the sixth-worst air quality in 2021[1]. The primary factor contributing to an increase in air pollutants is the unchecked use of motorized vehicles and various industries in daily living [2].

The primary factor causing natural change is air pollution. The most impacted item is air, due to a mixture of dust particles with diameters ranging from 0.1 to 100 m. Because industrial waste will use clean air to create polluted air, this is especially concerning. PM10 dust particles are solid, airborne particles with a width of 10 micrometers (particulate matter). In contrast, indoors, dust with a particle size of less than 2.5 microns (PM2.5) is more prevalent and is thought to be a risk factor for respiratory illnesses. The two solid particles' nature allows them to settle so that they can damage respiratory tissues like the alveoli and bronchitis [3].In essence, nose hair serves as the first filter in the process of human respiration, and it is believed to be 50% to 60% effective at filtering dust particles larger than 10 microns (PM10) and having a diameter of up to 45 microns (total suspended particulate). This does not apply to PM2.5, a type of dust that cannot be filtered in the upper respiratory system and has a significant negative effect on the airways [4].

Environmental degradation and technological development are correlated. Especially when it comes to air pollution because of the way that pollutants suspended in the air directly affects people [5], [6]. Because they are directly released into the atmosphere, such as the results of forest fires, car exhaust, industrial, and other pollutant activities, these particles will more readily contribute to the rise in pollutant numbers. High amounts of pollutants often make lung cancer, cardiovascular disease, and respiratory conditions worse. Recycling is possible in a natural process if the concentration rate of the refuse is still proportional to the rate of the recycling process, allowing for the mitigation measure of control.According to research by Ref. [7], over the past 30 years, excessive fuel use, fossil fuel burning, global warming, and industrial and factory waste have been the major causes of air pollution, including in agriculture. Due to the release of aerosols that cling directly to the air in the atmosphere, this may result in decreased visibility. Therefore, it is crucial for this research to make accurate predictions in order to play a part in preventing air pollution, which can be used to manage air quality [8].

Numerous earlier investigations were conducted using statistical and physical predictions. Because statistical forecasting projects future occurrences, it concentrates more on historical data [9]. Comparatively to the physical prediction model, which is dependent on actual conditions like the atmospheric conditions that serve as the foundation for the pollutant diffusion method in cases involving aerodynamics, the statistical prediction model will take into account the presence of quantitative data [10]. The extensive distribution of pollutants is calculated using the physical forecasting method, which places more stress on numerical models as a reference. According to research by Ref. [11], statistical forecasting models, as opposed to physical forecasting models, can avoid very complicated methods, allowing for calculations that are relatively quick, inexpensive, and of high accuracy.This research tries to use machine learning and the Decision Tree (DT) algorithm to implement a statistical forecasting model. An essential component of the advancement of artificial intelligence (AI) technology is machine learning [12], [13]. DT is a decision-making program with conditional control statements that is structured like a decision tree. Although this model tends to disregard spatial influences and outside factors when used, which can lead to over fitting if an incorrect index is used in the preparation process, it is effective at addressing the challenges of

adjusting data and non-linear relationships [14].

II. METHODS

This research makes use of six different real-time air quality sensors that were acquired from satellites. The data preparation for the analysis in this study includes preparing the data in.csv format. The data is then prepared to avoid data that does not provide information by checking for missing data using the imputation technique. The analysis in this study employs the Decision Tree (DT) algorithm. Regarding the mathematical processing of the complete program, the design of the software that is created is crucial. The core of this software design is the creation of a system workflow that classifies data based on sensor readings using the DT technique.To predict air quality from sensor readings, the software design for this monitoring system involves creating a DT classification model. Two classification parameters—normal and containing pollutants—are used to group the classification findings. Fig. 1 depicts the software design pipeline for DT-based predictive system modelling.

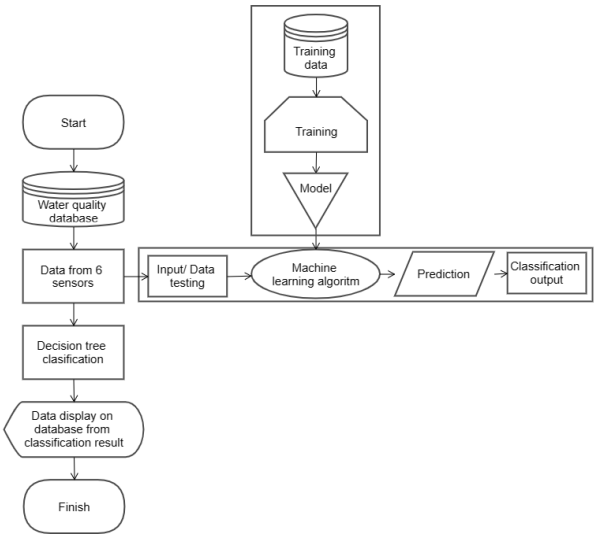

Fig. 1. The design flow of the prediction system model using DT.

This study utilizes 4895 data records in total. Table I presents the first five sensor data points from the 4895 data in more depth.

III. RESULTSAND DISCUSSION

One of the effects of harmful substances entering the sky or the atmosphere itself is air pollution. According to Government Regulation No. 41 of 1999, air pollutants include ozone (O3), carbon monoxide (CO), nitrogen dioxide (NO2), sulphur dioxide, and particulate matter measuring 10 microns (PM10) (SO2).

TABLE I
RAW DATA FROM SENSOR RECORDS

Id	Sensor					
	A	**B**	**C**	**D**	**E**	**F**
0	174	68	73	44	48	202
1	170	67	94	44	48	198
2	171	67	98	44	49	197
3	171	68	91	44	51	196
4	174	67	65	47	51	196

Carbon monoxide (CO), nitrogen dioxide (NO2), chlorofluorocarbons (CFC), sulphur dioxide (SO2), hydrocarbons (HC), particulates, tin (Pb), and carbon dioxide (CO2) are just a few of the potentially dangerous substances that can reach the atmosphere [15]. The most significant greenhouse gases in the atmosphere are carbon dioxide (CO2) and methane (CH4), which have a significant effect on radiation forces and the climate of the planet[16]. In landfills, the gases with the greatest concentrations are methane (CH4) and carbon dioxide (CO2) (TPA). Humans who inhale this gas may experience symptoms of respiratory issues like coughing, chest discomfort, and shortness of breath. Natural factors (such as smoke from fires or volcanic eruptions) or human actions can both contribute to air pollution (transportation, garbage disposal, industry). The main influence on air pollution is human activity, particularly the use of motorized transportation. According to a study, the biggest air pollutants in the world, including 44% TSP, 89% hydrocarbons, 100% Pb, and 73% NOx, are created by engine exhaust. These pollutants are harmful to human health, particularly for expectant mothers, young children, and parents [17]. Additionally, the effects of brief exposure to different air pollutants act as migraine and headache causes [18].

Data for comparison of N20, CH4, CO2, and CO2eq are presented in a comparison chart between 4 provinces in Indonesia, namely East Nusa Tenggara, DKI Jakarta, South Sulawesi, and East Java. This data is taken into consideration in light of existing pollution levels. DKI Jakarta as the capital city, South Sulawesi, and East Java are some of the most densely populated areas. East Nusa Tenggara is considered an area that is still relatively low in pollution levels. These data are in accordance with the results of research conducted by Ref. [19]. The comparison data completely shown in Fig. 2.

Fig. 2 demonstrates that DKI Jakarta has a very high pollution level when compared to the other 3 districts. In order to serve as a warning in the future, this represents a new challenge for this study. To avoid missing data, the data cleaning procedure is then performed [20], [21]. By focusing on the least squares alternative or optimizing the probability function, this method employs imputation [22]. Fig. 3 displays the imputation findings.

According to Fig. 3, the imputation method was effectively applied as evidenced by the color blocks on each feature, with black being the dominant color. This indicates that there are no outliers for any of the attributes. so that the model training procedure will be simplified by this process. Additionally, correlation analysis is used to ascertain the connection between sensor data.

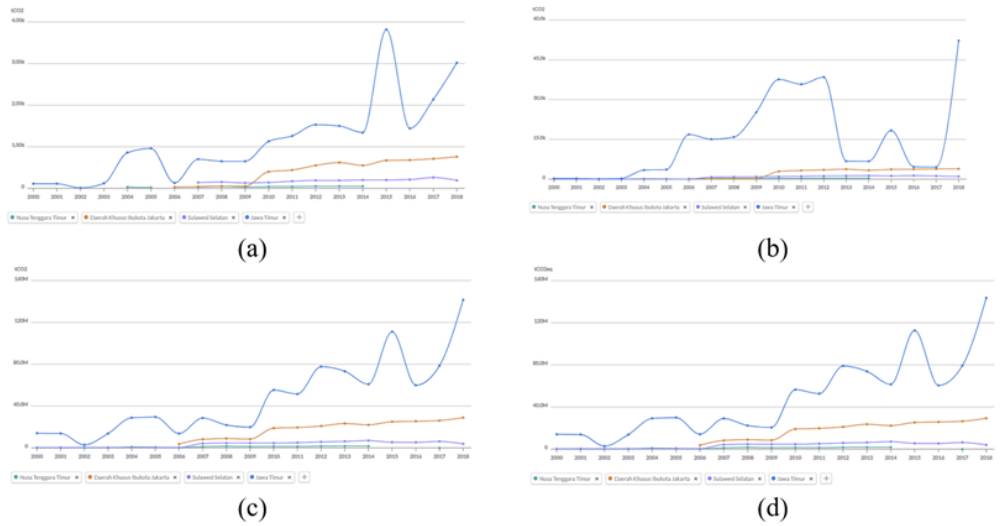

Fig. 2. Comparison graph of pollution levels (a) N20, (b) CH4, (c) CO2, and (d) CO2eq.

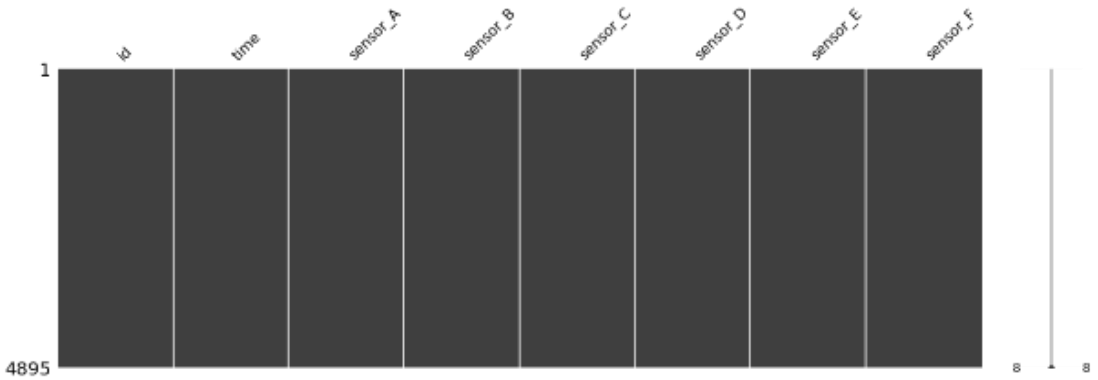

Fig. 3. Data cleansing.

According to these findings, a high degree of correlation existed, as evidenced by the color degradation, which ranged from 0.6 to 0.8. This shows that data normalization or data reduction are not necessary for the procedure as shown in Fig. 4.

The percentage of sensor variance is described as a function of the number of clusters [23]. The first cluster will provide a lot of information about the angle effect, so that it forms an angle as shown in Fig. 5. This is in accordance with the data plots generated for each attribute.

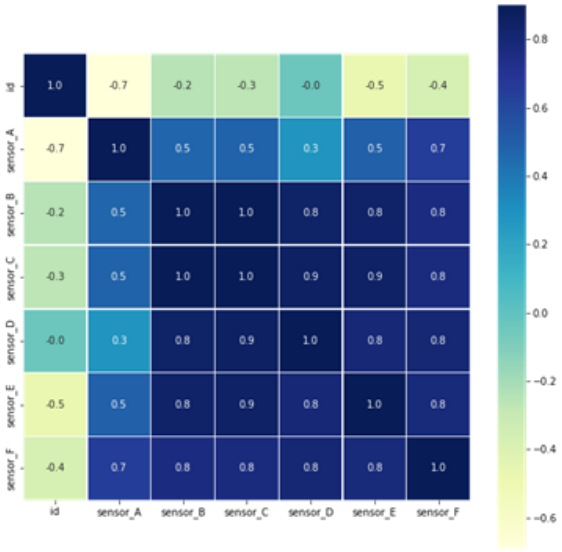

Fig. 4. Correlation of sensor data heat maps.

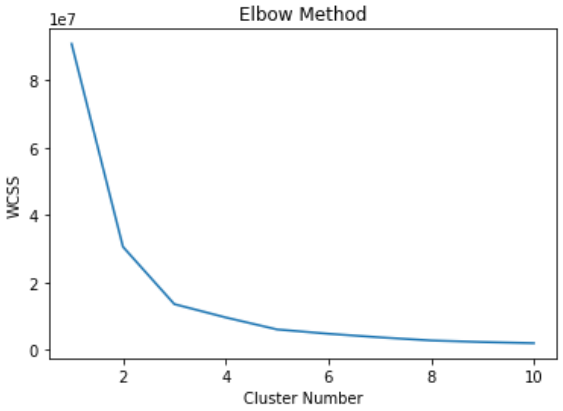

Fig. 5. Elbow method.

The number of clusters obtained based on data fractures with the elbow method[24] has two elbow fractures, so this number is the result of optimal cluster formation [25]. So that the cluster data output is shown in Table II.

TABLE II
CLUSTERED DATA

Sensor						Klaster
A	B	C	D	E	F	
174	68	73	44	48	202	2
170	67	94	44	48	198	2
171	67	98	44	49	197	2
171	68	91	44	51	196	2
174	67	65	47	51	196	2
…	…	…	…	…	…	…
56	62	43	130	36	133	2
56	64	45	131	38	133	2
55	70	49	132	42	132	2
55	71	50	133	43	132	2
56	64	44	131	38	133	2

The results of model training obtained a mean absolute error (MAE) of 0.84 and a residual sum of squares (MSE) of 1.50, while the value of the coefficient of determination (R^2) was obtained at 0.76. Complete accuracy and precision are presented in Fig. 6.

	precision	recall	f1-score	support
0	0.93	1.00	0.96	1010
1	1.00	0.93	0.96	14
2	0.99	0.83	0.90	445
accuracy			0.95	1469
macro avg	0.97	0.92	0.94	1469
weighted avg	0.95	0.95	0.95	1469

Fig. 6. Model training results.

Based on Fig. 6, it is found that the accuracy of using DT is 0.95. From these results, it is then followed by the process of develop a prediction system based on the model so that it is easier to use. The following display is seen in Fig. 7.

IV. CONCLUSION

The distribution of data obtained from the sensors recorded as many as 4895 records. This distribution is analysed based on the six attribute sensor data sets obtained from satellite data. The results of the analysis obtained a level of correlation between attributes according to colour degradation recorded in the range of 0.6 to 0.8, which can be categorized as high.

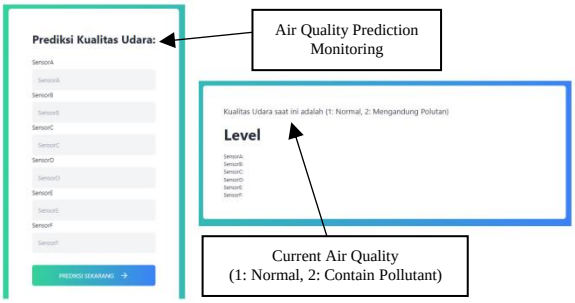

Fig. 7. Results of prediction system development.

This study also presents the Elbow Method, which shows an appropriate optimization value for the trial and error values, which is 2. So that the process of determining clusters based on the results of data classification takes as many as 2. From these results, it is followed by a statistical forecasting process using the DT algorithm, which obtains an accuracy of 0.95. These results are also reinforced by several statistical calculations from model training, namely the mean absolute error (MAE) obtained at 0.84 and the residual sum of squares (MSE) at 1.50. While the value of the coefficient of determination (R^2) is 0.76. From these results, a model was developed to obtain an attractive interface application that could be used as a warning system application for monitoring air quality in the future.

REFERENCES

[1] A. Amalia, A. Zaidiah, and I. N. Isnainiyah, "Prediksikualitasudaramenggunakanalgoritma k- nearest neighbor," *JIPI (JurnalIlmiahPenelitiandanPembelajaranInformatika)*, vol. 7, no. 2, pp. 496–507, 2022.

[2] M. A. Fath, "Literature Review : PengaruhKualitasUdaradanKondisiIklimTerhadapPerekonomianMasyarakat," *Media GiziKesmas*, vol. 10, no. 2, pp. 329–342, 2021.

[3] T. He, L. Jin, and X. Li, "On the triad of air PM pollution, pathogenic bioaerosols, and lower respiratory infection," *Environ Geochem Health*, vol. 7, 2021, doi: 10.1007/s10653-021-01025-7.

[4] S. Jeong*et al.*, "PM2.5 Exposure in the Respiratory System Induces Distinct Inflammatory Signaling in the Lung and the Liver of Mice," *J Immunol Res*, vol. 2019, 2019, doi: 10.1155/2019/3486841.

[5] U. A. Hvidtfeldt*et al.*, "Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon," *Environmental Epidemiology*, vol. 2, no. 2, p. e014, 2018, doi: 10.1097/ee9.0000000000000014.

[6] Y. Gonzalez *et al.*, *Inhaled Air Pollution Particulate Matter In Alveolar Macrophages Alters Local Pro-Inflammatory Cytokine And Peripheral IFNγ Production In Response To Mycobacterium Tuberculosis*. American Thoracic Society, 2017.

[7] S. Neelakandan, M. A. Berlin, S. Tripathi, V. B. Devi, I. Bhardwaj, and N. Arulkumar, "IoT-based traffic prediction and traffic signal control system for smart city," *Soft comput*, vol. 25, no. 18, pp. 12241–12248, 2021, doi: 10.1007/s00500-021-05896-x.

[8] D. Saravanan, D. K. S. Kumar, R. Sathya, and U. Palani, "An Iot based air quality monitoring and air pollutant level prediction system using machine learning approach–Dlmnn," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 4, pp. 925–945, 2020.

[9] S. Neelakandan and D. Paulraj, "An automated exploring and learning model for data prediction using balanced CA-SVM," *J Ambient IntellHumanizComput*, vol. 12, no. 5, pp. 4979–4990, 2021, doi: 10.1007/s12652-020-01937-9.

[10] Q. Xiao, H. H. Chang, G. Geng, and Y. Liu, "An Ensemble Machine-Learning Model to Predict Historical PM2.5 Concentrations in China from Satellite Data," *Environ SciTechnol*, vol. 52, no. 22, pp. 13260–13269, 2018, doi: 10.1021/acs.est.8b02917.

[11] Y. A. Aliyu and J. O. Botai, "Appraising city-scale pollution monitoring capabilities of multi-satellite datasets using portable

pollutant monitors," *Atmos Environ*, vol. 179, no. November 2017, pp. 239–249, 2018, doi: 10.1016/j.atmosenv.2018.02.034.
[12] S. Saputra, A. Yudhana, and R. Umar, "Implementation of Naïve Bayes for Fish Freshness Identification Based on Image Processing," *JURNAL RESTI (RekayasaSistemdanTeknologiInformasi)*, vol. 6, no. 3, pp. 412–420, 2022, doi: https://doi.org/10.29207/resti.v6i3.4062.
[13] A. Yudhana, R. Umar, and S. Saputra, "Fish Freshness Identification Using Machine Learning: Performance Comparison of k-NN and Naïve Bayes Classifier," *Journal of Computing Science and Engineering*, vol. 16, no. 3, pp. 153–164, Sep. 2022, doi: 10.5626/JCSE.2022.16.3.153.
[14] J. Amanollahi and S. Ausati, "Validation of linear, nonlinear, and hybrid models for predicting particulate matter concentration in Tehran, Iran," *TheorApplClimatol*, vol. 140, no. 1–2, pp. 709–717, 2020, doi: 10.1007/s00704-020-03115-5.
[15] M. Dwangga, "IntensitasPolusiUdaraUntukPenunjangPenataanRuang Kota PelaihariKabupaten Tanah Laut," *MetodeJurnalTeknikIndustri*, vol. 4, no. 2, pp. 69–77, 2018.
[16] M. Li *et al.*, "Human metabolic emissions of carbon dioxide and methane and their implications for carbon emissions," *Science of the Total Environment*, vol. 833, no. April, 2022, doi: 10.1016/j.scitotenv.2022.155241.
[17] S. T. Fandani, H. Sulistiyowati, and R. Setiawan, "Tingkat PencemaranUdara di Desa Silo dan Pace, Kecamatan Silo, KabupatenJemberdenganMenggunakan Lichen SebagaiBioindikator," *BerkalaSainstek*, vol. 7, no. 2, p. 39, 2019, doi: 10.19184/bst.v7i2.6861.
[18] H. Elser*et al.*, "Correction to: Air pollution, methane super-emitters, and oil and gas wells in Northern California: the relationship with migraine headache prevalence and exacerbation (Environmental Health, (2021), 20, 1, (45), 10.1186/s12940-021-00727-w)," *Environ Health*, vol. 20, no. 1, pp. 1–14, 2021, doi: 10.1186/s12940-021-00745-8.
[19] A. N. Alifah, H. N. Fadhilah, and T. M. Sianipar, "KlasterisasiKabupaten/Kota di Jawa Barat Berdasarkan Tingkat KenyamanandenganMetode K-Means Clustering," in *Seminar NasionalSains Data*, 2022, vol. 2022, pp. 30–38.
[20] E. Sartika, "AnalisisMetode K Nearest Neighbor Imputation (KNNI) untukMengatasi Data HilangPadaEstimasi Data Survey," *Jurnal TEDC*, vol. 12, no. 3, pp. 219–227, 2018.
[21] M. R. Andryan, M. Fajri, and N. Sulistyowati, "KomparasiKinerjaAlgoritmaXgboostdanAlgoritma Support Vector Machine (SVM) untuk Diagnosis PenyakitKankerPayudara," *JIKO (JurnalInformatikadanKomputer)*, vol. 6, no. 1, pp. 1–5, Feb. 2022, doi: 10.26798/jiko.v6i1.500.
[22] A. Jufriansah, Y. Pramudya, A. Khusnani, and S. Saputra, "Analysis of Earthquake Activity in Indonesia by Clustering Method," *Journal of Physics: Theories and Applications*, vol. 5, no. 2, pp. 92–103, 2021, doi: 10.20961/jphystheor-appl.v5i2.59133.
[23] J. A. Prabowo and H. Dhika, "Safe Routing Model and Balanced Load Model for Wireless Sensor Network," *JurnalPendidikanTeknologiInformasi*, vol. 5, no. 1, pp. 44–58, 2021.
[24] R. Yuliana Sari, H. Oktavianto, and H. WahyuSulistyo, "Algoritma K-Means denganMetode Elbow untukMengelompokkanKabupaten/Kota Di Jawa Tengah BerdasarkanKomponenPembentukIndeks Pembangunan Manusia," *Jurnal Smart Teknologi*, vol. 3, no. 2, pp. 104–108, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST
[25] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int J ComputAppl*, vol. 105, no. 9, pp. 975–8887, 2014.

Adi Jufriansah is a lecture at the Physics Education Study Program, IKIP Muhammadiyah Maumere, Indonesia. His area of expertise on arificial intelegence. He has many publications ini various reputable journal. (email: saompu@gmail.com).

Azmi Khusnani is a lecture at the Physics Education Study Program, IKIP Muhammadiyah Maumere, Indonesia. Her research is interest about Experiment of Physics. She has many publication in various reputable journal. (email: husnaniazmi@gmail.com).

Yudhiakto Pramudya has Doctoral graduate in physics at Wesleyan University, USA in the field of superfluids. Currently working as a lecturer at Ahmad Dahlan University in Yogyakarta with research on vibrations and waves. (email: yudhiakto.pramudya@pfis.uad.ac.id).

Nursina Sya'bania is a lecture at the Chemistry Education Study Program, IKIP Muhammadiyah Maumere, Indonesia. She has a research interest in the learning media technology. She has many publications in various reputable journal. (email: nisa.syabania@gmail.com).

Kristina Tresia Leto is now teaching at Chemistry Education Study Program of IKIP Muhammadiyah Maumere. Her research interest is about Analytical Chemistry, Instrumentation and Chemistry Education. She loves reading, cooking and gardening. (email: kristinatresia922@gmail.com).

Hamzarudin Hikmatiar is one of the lecturers at IKIP Muhammadiyah Maumere and also part of the Center for Astronomy Studies (PUDIASTRO) IKIP Muhammadiyah Maumere. In addition, as the initiator of Langit Sikka who is involved in the world of education about the universe. (email: hamzarudinhikmatiar90@gmail.com).

Sabarudin Saputra is a student in Master Program of Informatics at Universitas Ahmad Dahlan, Indonesia. His research is focused on mathematics, image processing, and artificial intelligence. (email: sicoccinela@gmail.com).