Jurnal Publikasi Ilmu Komputer dan Multimedia (JUPIKOM)
Vol.2, No.2 Mei 2023
E-ISSN : 2808-8980 P-ISSN :2808-9383, Hal 86-92
DOI: https://doi.org/10.55606/jupikom.v2i2.18578

Klasifikasi Pencemaran Udara Di DKI Jakarta Menggunakan Metode Naïve Bayes

Hilda Rachmi
Universitas Bina Sarana Informatika

Alamat: Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Jakarta Pusat
Korespondensi penulis: hilda.hlr@bsi.ac.id

Abstract. *Air quality index in Indonesia shows Jakarta in 5th position with the highest air pollution. Naïve Bayes is one of the data mining methods that has excellent accuracy with simple calculations in the field of classification learning. In this study, Naïve Bayes is used to classify air quality based on air pollution standard index data in the Special Capital Region of Jakarta within 12 (twelve) months. The processed variables include PM10, PM25, SO2, CO, O3, and NO2. By using accuracy value evaluation data, the results show that Naïve Bayes is effectively used for air pollution classification with an accuracy rate of 91,96%..*

Keywords:*Air Pollution, Clasiffication, Naïve Bayes*.

Abstrak. Perhitungan indeks kualitas udara pada kota di Indonesia menunjukkan Jakarta berada pada posisi kelima dengan polusi udara tertinggi. Naïve Bayes sebagai salah satu metode pada data mining yang memiliki akurasi sangat baik dengan perhitungan sederhana dalam bidang pembelajaran klasifikasi. Pada penelitian ini Naïve Bayes digunakan untuk klasifikasi kualitas udara berdasarkan data indeks standar pencemaran udara di Daerah Khusus Ibukota Jakarta dalam rentang waktu 12 (dua belas) bulan. Variabel yang diolah meliputi PM10, PM25, SO2, CO, O3, dan NO2. Dengan menggunakan data evaluasi nilai akurasi hasil penelitian menunjukkan Naïve Bayes efektif digunakan untuk klasifikasi pencemaran udara dengan tingkat akurasi sebesar 91,96%.

Kata kunci: Klasifikasi, Naïve Bayes, Pencemaran Udara.

LATAR BELAKANG

Pada bulan Agustus 2023 data Indeks kualitas udara (AQI) menunjukkan tingkat polusi udara di Jakarta tidak sehat dengan besaran indeks 154. Konsentrasi polutan utama atau $PM_{2.5}$ di Jakarta saat ini 12.4 kali lebih besar dari nilai panduan kualitas udara tahunan WHO. Tingkat polusi udara ini diperkirakan menyebabkan 8.400 kematian di Jakarta dengan kerugian sebesar $2.200.000.000 USD (*Kualitas Udara Di Jakarta*, n.d.). Permasalahan pencemaran udara di DKI Jakarta dapat dikatakan paling besar dan menantang bagi pemerintah (Bernadet et al., 2023)

Pencegahan dan pengelolaan polusi sangat penting untuk dilakukan melalui tindakan pendeteksian, pengontrolan, dan evaluasi indeks pencemaran untuk mengetahui

Received September 30, 2022; Revised Januari 30, 2023; Accepted Mei 30, 2023
Hilda Rachmi, hilda.hlr@bsi.ac.id

bahwa indeks pencemaran sudah melebihi standar yang ditetapkan (Xu et al., 2023). Pengukuran Indeks Standar Pencemar Udara di Indonesia dilakukan berdasarkan peraturan menteri (Permen LHK Nomor 14 Tahun 2020, 2020). Prediksi kualitas udara menjadi penting karena berhubungan dengan kesehatan (Sadhasivam et al., 2021). Dengan memprediksi kualitas udara dapat membantu pemerintah dan organisasi lingkungan untuk menerapkan langkah-langkah pencegahan, termasuk mencegah kematian akibat polusi udara (Londhe, 2021).

Beberapa tahun terakhir proses pengumpulan, pengelolaan, dan analisis data berkembang ke arah baru untuk memahami pengetahuan. Penelitian mengenai kualitas udara juga banyak dilakukan dengan menggunakan berbagai sumber data untuk mendapatkan gambaran yang lebih detail (Represa et al., 2019). Artikel yang mempelajari kualitas udara dengan perspektif yang berbeda memberikan ringkasan dari berbagai alat yang tersedia.

KAJIAN TEORITIS

Penelitian mengenai polusi udara dengan menggunakan analisis time series serta model Autoregressive Integrated Moving Average model (ARIMA) menunjukkan ARIMA lebih akurat dari metode Prophet pada beberapa faktor (Ye, 2019). Pendekatan data mining dengan K-NN digunakan pada klasifikasi indikator tingkat kualitas udara di Kota Palembang (Sanmorino et al., 2022) dan Makasar (Aini & Mustafa, 2020). Modifikasi algoritma untuk mengoptimalkan parameter dengan menggunakan dataset dari 13 kota di Beijing memperoleh data yang efisien dan andal untuk pemantauan kualitas udara (Huang et al., 2021). Data Inventarisasi Polutan Nasional Australia (NPI) selama tahun 2008–2018 dianalisa untuk memperkirakan polutan udara (Hendryx et al., 2020). Prediksi polutan udara juga dilakukan melalui eksplorasi interaksi polutan pada 3 kota di Iran dengan menggunakan tiga kriteria penilaian yang berbeda untuk menilai model XGBoost sebagai bentuk pengembangan model berbasis pembelajaran mesin (Rad et al., 2022).

Penelitian mengenai polusi udara di Jakarta sudah dilakukan dengan beberapa metode, seperti: LSTM dan GRU (Handhayani, 2023), K-NN (Wiranata et al., 2023)(Amalia et al., 2022), Linear Regression Model (Latief & Karyanti, 2022), Metode C.45 (Astriyani et al., 2023) dan (Umri et al., 2021), K-Means (Sitorus et al., 2022), serta komparasi antara Neural Network, Support Vector Machine, K-Nearest Neighbors dan Decission Tree (Nurdalia et al., 2023).

Pada penelitian ini penulis menyajikan hasil pengolahan data dengan menggunakan tools data mining dan data Indeks Standar Pencemar Udara (ISPU) dari Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia. Data diambil dari 5 stasiun pemantau kualitas udara (SPKU): DKI1 (Bunderan HI), DKI2 (Kelapa Gading), DKI3 (Jagakarsa), DKI4 (Lubang Buaya), dan DKI5 (Kebon Jeruk).

METODE PENELITIAN

Penelitian ini merupakan penelitian bidang data mining yang merupakan kegiatan untuk mencari pola dari data yang berjumlah besar yang tersimpan dalam database, data warehouse, atau penyimpanan informasi lainnya (Susana et al., 2022). Metode yang

digunakan adalah Naïve Bayes. Pengklasifikasi Bayesian merupakan pengklasifikasi factual yang dapat mengantisipasi probabilitas partisipasi. Posisi yang menguntungkan dari pengklasifikasi Bayes adalah bahwa dibutuhkan sejumlah informasi persiapan yang terbatas untuk mengevaluasi batas-batas (cara dan fluktuasi faktor) yang penting untuk pengaturan (Vanakovarayan et al., 2020). Naïve Bayes menggunakan perhitungan probabilitas dan konsep dasar Teorema Bayes, yaitu dengan melakukan klasifikasi melalui perhitungan nilai probabilitas (Susana et al., 2022).

Penerapan data mining pada penelitian ini menggunakan proses tahapan knowledge discovery in databases dengan tahapan sebagai berikut:

1. *Data Collecting*

Data yang diperlukan sebelum dilakukan tahap penggalian informasi dalam *Knowledge Discovery Database* (KDD) dimulai dengan pengumpulan data. Data diambil dalam bentuk *Comma Separated Values* yang berjumlah 1825 data.

2. *Data Cleaning*

Proses data cleansing bertujuan untuk menghilangkan data yang tidak memiliki nilai (null), data yang salah input, data yang tidak relevan, duplikat data dan data yang tidak konsisten. Dari 1826 data yang telah dikumpulkan pada tahapan sebelumnya, didapatkan hasil akhir 1517 data setelah dibersihkan.

3. *Data Transformation*

Data transformation dilakukan dengan memberikan inisialisasi terhadap data. Pada tahapan ini kumpulan data dalam bentuk Comma Separated Values ditransformasi menjadi .xls agar dapat diproses menggunakan tools data mining.

4. *Data Mining*.

Pada tahap ini dilakukan penerapan metode dengan menggunakan Naïve Bayes.

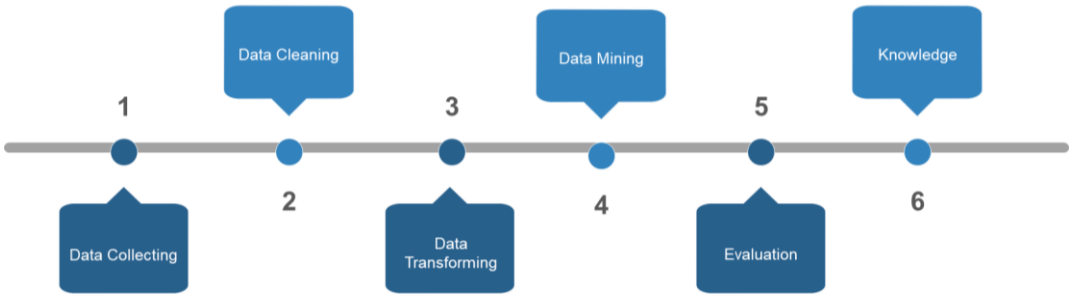
5. *Evaluation*

Tahap evaluasi dilakukan dengan melihat kinerja metode yang dapat dilihat melalui tingkat akurasi.

6. *Knowledge*

Tahap ini dilakukan untuk membahas metode yang digunakan dan menentukan apakah metode tersebut bekerja dengan baik.

Tahapan penelitian seperti yang dijelaskan di atas dapat dilihat pada Gambar 1



Gambar 1. Tahapan Penelitian

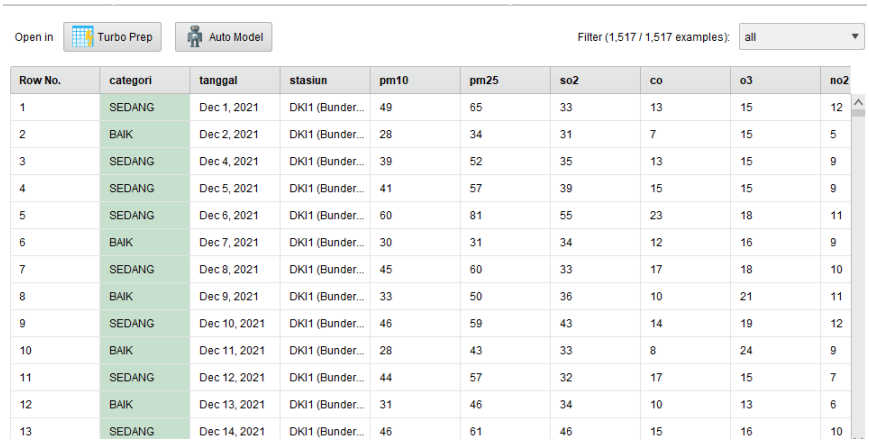
HASIL DAN PEMBAHASAN

Selama rentang waktu 12 (dua belas) bulan dikumpulkan data sebanyak 1826 yang mencakup data tanggal pengambilan, stasiun pengambilan data, variabel yang dipantau oleh stasiun pemantau udara meliputi Partikulat PM10 dan PM25, Sulfida SO2, Carbon Monoksida CO, Ozon O3, dan Nitrogen Dioksida NO2, variabel Max menunjukkan nilai ukur paling tinggi dari seluruh parameter dalam waktu pengukuran yang sama serta variabel Critical menunjukkan Parameter yang hasil pengukurannya paling tinggi. Detail data Indeks Standar Pencemar Udara Tahun 2021 dapat dilihat pada Tabel 1.

Tabel 1. Data Indeks Standar Pencemar Udara Tahun 2021

No	Tanggal	Stasiun	PM10	PM25	SO2	CO	O3	NO2	Max	Critical
1	01/01/2021	DKI1 (Bunderan HI)	38	53	29	6	31	13	53	PM25
2	02/01/2021	DKI1 (Bunderan HI)	27	46	27	7	47	7	47	O3
3	03/01/2021	DKI1 (Bunderan HI)	44	58	25	7	40	13	58	PM25
4	04/01/2021	DKI1 (Bunderan HI)	30	48	24	4	32	7	48	PM25
5	05/01/2021	DKI1 (Bunderan HI)	38	53	24	6	31	9	53	PM25
….	……..	……..	….	….	….	….	….	….	….	……..
1824	29/12/2021	DKI5 (Kebon Jeruk) Jakarta Barat	34	54	28	8	25	29	54	PM25
1825	30/12/2021	DKI5 (Kebon Jeruk) Jakarta Barat	53	75	25	15	23	44	75	PM25
1826	31/12/2021	DKI5 (Kebon Jeruk) Jakarta Barat	60	87	28	19	30	53	87	PM25

Data yang dikumpulkan kemudian dibersihkan dengan menggunakan blok Replace Missing Value. Proses ini menghasilkan data bersih yang tidak memiliki nilai null pada setiap variabel. Hasil dari proses data cleansing ditampilkan pada gambar 2.

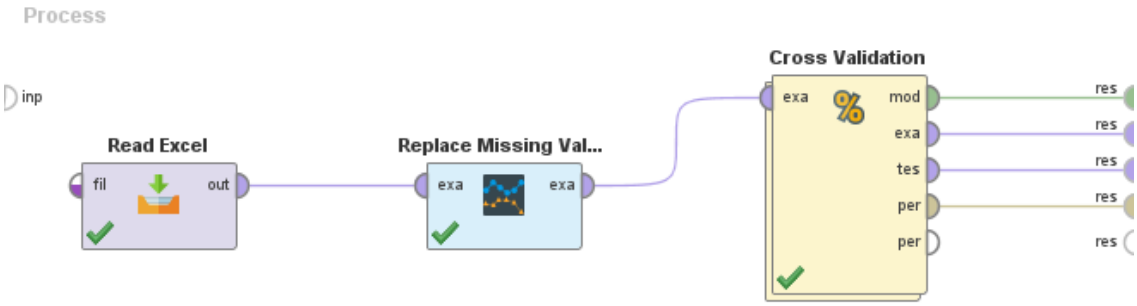

Open in Turbo Prep Auto Model Filter (1,517 / 1,517 examples): all

Row No.	categori	tanggal	stasiun	pm10	pm25	so2	co	o3	no2
1	SEDANG	Dec 1, 2021	DKI1 (Bunder...	49	65	33	13	15	12
2	BAIK	Dec 2, 2021	DKI1 (Bunder...	28	34	31	7	15	5
3	SEDANG	Dec 4, 2021	DKI1 (Bunder...	39	52	35	13	15	9
4	SEDANG	Dec 5, 2021	DKI1 (Bunder...	41	57	39	15	15	9
5	SEDANG	Dec 6, 2021	DKI1 (Bunder...	60	81	55	23	18	11
6	BAIK	Dec 7, 2021	DKI1 (Bunder...	30	31	34	12	16	9
7	SEDANG	Dec 8, 2021	DKI1 (Bunder...	45	60	33	17	18	10
8	BAIK	Dec 9, 2021	DKI1 (Bunder...	33	50	36	10	21	11
9	SEDANG	Dec 10, 2021	DKI1 (Bunder...	46	59	43	14	19	12
10	BAIK	Dec 11, 2021	DKI1 (Bunder...	28	43	33	8	24	9
11	SEDANG	Dec 12, 2021	DKI1 (Bunder...	44	57	32	17	15	7
12	BAIK	Dec 13, 2021	DKI1 (Bunder...	31	46	34	10	13	6
13	SEDANG	Dec 14, 2021	DKI1 (Bunder...	46	61	46	15	16	10

Gambar 1. Hasil Data Cleansing

Proses validasi setelah pembacaan dan pembersihan data file data menggunakan Cross Validation dengan 10 folds yang bertujuan mengukur keakuratan kinerja model yang digunakan. Data yang didapat pada proses cleansing dipecah menjadi beberapa

subset dengan ukuran yang sama. Selanjutnya satu subset akan dipertahankan sebagai data pengujian dan sisanya akan digunakan sebagai data pelatihan yang divalidasi secara berulang sebanyak k folds. Subproses pada operator validasi ini berupa testing dan training dengan rasio 0,7 dan 0,3.



Gambar 2. Proses Cross Validation

Proses selanjutnya adalah memasukan data ke dalam proses pemodelan Naïve Bayes dengan operator apply model serta operator performance utnuk mendapatkan hasil akhir dalam bentuk pengukuran akurasi. Hasil pengukuran akurasi yang diperoleh dari sebesar 91,96%. Data yang diproses akan diklasifikasi menjadi 3 class yaitu, Sedang, Baik, dan Tidak Sehat. Jumlah class Sedang yang benar diklasifikasikan ke dalam class Sedang oleh classifier yaitu 1030 data dan jumlah class Sedang yang diklasifikasikan ke dalam class Tidak Sehat oleh classifier yaitu 5 data dengan class precission sebesar 99.52%. Sedangkan class Baik yang diklasifikasikan ke dalam class Sedang oleh classifier yaitu 63 data dan jumlah class Baik yang diklasifikasikan ke dalam class Baik oleh classifier yaitu 125 data dengan class precission sebesar 66,49%. Jumlah class Tidak Sehat yang diklasifikasikan ke dalam class Sedang oleh classifier yaitu 54 data dan jumlah class Tidak Sehat yang benar diklasifikasikan ke dalam class Tidak Sehat oleh classifier yaitu 240 data dengan class precission sebesar 81,83%. Untuk class recall dengan klasifikasi benar class Sedang sebesar 89,80%, class recall dengan klasifikasi benar class Baik sebesar 100%, dan class recall dengan klasifikasi benar class Tidak Sehat sebesar 97,96%. Tabel akurasi untuk penelitian ini dapat dilihat pada gambar 3.

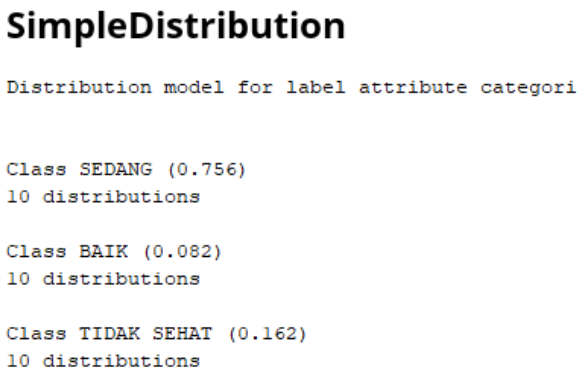
accuracy: 91.96% +/- 1.96% (micro average: 91.96%)

	true SEDANG	true BAIK	true TIDAK SEHAT	class precision
pred. SEDANG	1030	0	5	99.52%
pred. BAIK	63	125	0	66.49%
pred. TIDAK SEHAT	54	0	240	81.63%
class recall	89.80%	100.00%	97.96%	

Gambar 3. Hasil Akurasi

Dari hasil pengujian model terbentuk model distribusi sebanyak 3 class yang masing-masing memiliki 10 class distribusi. Nilai class Sedang sebesar 0,756%, nilai class Baik

sebesar 0,082% dan nilai class Tidak Sehat seebsar 0,162% seperti yang ditunjukkan pada gambar 4.



Gambar 4. Sebaran Distribusi Class

KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian dalam penilaian akurasi data Indeks Standar Pencemar Udara Tahun 2021 menggunakan algoritma Naïve Bayes, variable data yang digunakan yaitu Partikulat PM10 dan PM25, Sulfida SO2, Carbon Monoksida CO, Ozon O3, dan Nitrogen Dioksida NO2, Max, dan Critical. Metode Naïve bayes dapat diimplementasikan dengan baik serta dapat mengidentifikasi class kualitas udara dengan akurasi sebesar 91,96% berdasarkan 3 class yang terbentuk, yaitu Sedang, Baik, dan Tidak Sehat.

DAFTAR REFERENSI

Aini, N., & Mustafa, M. S. (2020). Data Mining Approach to Predict Air Pollution in Makassar. *2020 2nd International Conference on Cybernetics and Intelligent System, ICORIS 2020*, *1*. https://doi.org/10.1109/ICORIS50180.2020.9320800

Amalia, A., Zaidiah, A., & Isnainiyah, I. N. (2022). PREDIKSI KUALITAS UDARA MENGGUNAKAN ALGORITMA K- NEAREST NEIGHBOR. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika) Volume*, *7*(2), 98–104. https://doi.org/10.33387/jiko.v4i2.2871

Astriyani, M., Laela, I. N., Lestari, D. P., Anggraeni, L., & Astuti, T. (2023). Analisis Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma C.45. *JuSiTik : Jurnal Sistem Dan Teknologi Informasi Komunikasi*, *6*(1), 36–41. https://doi.org/10.32524/jusitik.v6i1.790

Bernadet, Listyarini, S., & Warlina, L. (2023). Pengaruh Kebijakan Pencemaran Udara Sektor Transportasi Terhadap Nilai Indeks Kualitas Udara (Iku) Di Dki Jakarta. *JURNAL PENDIDIKAN LINGKUNGAN DAN PEMBANGUNAN BERKELANJUTAN*, *24*(01), 1–13. https://doi.org/10.21009/plpb.v24i01.30798

Handhayani, T. (2023). An integrated analysis of air pollution and meteorological

conditions in Jakarta. *Scientific Reports*, *13*(1), 1–11. https://doi.org/10.1038/s41598-023-32817-9

Hendryx, M., Islam, M. S., Dong, G. H., & Paul, G. (2020). Air Pollution Emissions 2008–2018 from Australian Coal Mining: Implications for Public and Occupational Health. *International Journal of Environmental Research and Public Health*, *17*(5). https://doi.org/10.3390/ijerph17051570

Huang, Y., Deng, Y., Wang, C., & Fu, T. (2021). Hybrid Data Mining Forecasting System Based on Multi-Objective Optimization and Selection Model for Air pollutants. *Frontiers in Environmental Science*, *9*(December). https://doi.org/10.3389/fenvs.2021.761287

Kualitas udara di Jakarta. (n.d.). https://www.iqair.com/id/indonesia/jakarta

Latief, M. A., & Karyanti, Y. (2022). Data Mining & Analytic Forecasting Indeks Standar Pencemar Udara Jakarta Menggunakan Metode Linear Regression (Studi Kasus: Dataset Indeks Standar Pencemar Udara Jakarta 2021). *Journal Of Social Research*, *1*(10), 1164–1176. https://doi.org/10.55324/josr.v1i10.248

Londhe, M. M. (2021). Data Mining and Machine Learning Approach for Air Quality Index Prediction. *International Journal of Engineering and Applied Physics (IJEAP)*, *1*(2), 136–153. https://ijeap.org/

Nurdalia, Zilrahmi, Permana, D., & Salma, A. (2023). Comparison of Naïve Bayes and K-Nearest Neighbor for DKI Jakarta Air Pollution Standard Index Classification. *UNP Journal of Statistics and Data Science*, *1*(2), 67–73. https://doi.org/10.24036/ujsds/vol1-iss2/29

Permen LHK Nomor 14 Tahun 2020, PERATURAN MENTERI LINGKUNGAN HIDUP DAN KEHUTANAN REPUBLIK INDONESIA NOMOR P.14/MENLHK/SETJEN/KUM.1/7/2020 TENTANG INDEKS STANDAR PENCEMAR UDARA 1 (2020).

Rad, A. K., Shamshiri, R. R., Naghipour, A., Razmi, S. O., Shariati, M., Golkar, F., & Balasundram, S. K. (2022). Machine Learning for Determining Interactions between Air Pollutants and Environmental Parameters in Three Cities of Iran. *Sustainability (Switzerland)*, *14*(13). https://doi.org/10.3390/su14138027

Represa, N. S., Fernández-Sarría, A., Porta, A., & Palomar-Vázquez, J. (2019). Data Mining Paradigm in the Study of Air Quality. *Environmental Processes*. https://doi.org/10.1007/s40710-019-00407-5

Sadhasivam, J., Muthukumaran, V., Thimmia Raja, J., Vinothkumar, V., Deepa, R., & Nivedita, V. (2021). Applying data mining technique to predict trends in air pollution in Mumbai. *Journal of Physics: Conference Series*, *1964*(4). https://doi.org/10.1088/1742-6596/1964/4/042055

Sanmorino, A., Alie, J., Ariati, N., & Wulanda, S. V. (2022). K-NN Based Air Classification as Indicator of the Index of Air Quality in Palembang. *SinkrOn*, *7*(3), 853–859. https://doi.org/10.33395/sinkron.v7i3.11469

Sitorus, M., Fitron, D., & Wisesa, C. A. S. (2022). Implementasi Algoritma K-Means Menggunakan Aplikasi Orange dalam Clustering Pencemaran Udara di DKI Jakarta Tahun 2021. *Journal of Informatics and Advanced Computing (JIAC)*, *3*(2), 161–164.

Susana, H., Suarna, N., Fathurrohman, & Kaslani. (2022). Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. *Jurnal Riset Sistem Informasi Dan Teknologi Informasi (JURSISTEKNI)*, *4*(1), 1–8. https://doi.org/10.52005/jursistekni.v4i1.96

Umri, S. S. A., Firdaus, M. S., & Primajaya, A. (2021). ANALISIS DAN KOMPARASI ALGORITMA KLASIFIKASI DALAM INDEKS PENCEMARAN UDARA DI DKI JAKARTA. *JIKO (Jurnal Informatika Dan Komputer)*, *4*(2), 98–104. https://doi.org/10.33387/jiko

Vanakovarayan, S., Murali, D., Prasanna, S., & Priyaradhikadevi, T. (2020). *J48 , CART AND NAVIE BAYESIAN ALGORTHIM FOR PERFORMANCE ANALYSIS OF SOFTWARE*. *7*(16), 2435–2440.

Wiranata, A. D., Soleman, Irwansyah, Sudaryana, I. K., & Rizal. (2023). 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.

Xu, Q., Ning, L., Yuan, T., & Wu, H. (2023). Application of data mining combined with power data in assessment and prevention of regional atmospheric pollution. *Energy Reports*, *9*, 3397–3405. https://doi.org/10.1016/j.egyr.2023.02.016

Ye, Z. (2019). Air Pollutants Prediction in Shenzhen Based on ARIMA and Prophet Method. *E3S Web of Conferences*, *136*, 1–5. https://doi.org/10.1051/e3sconf/201913605001