E-ISSN : 2722-0346

Implementasi Algoritma K-Means Menggunakan Aplikasi Orange dalam Clustering Pencemaran Udara di DKI Jakarta Tahun 2021

Michael Sitorus[1], Depriansa Fitron[2], Carolus Agung Segara Wisesa[3]
Program Studi Sistem Informasi
Institut Teknologi dan Bisnis Bank Rakyat Indonesia
michael.sitorus@cyber-univ.ac.id[1], defitronnex@gmail.com[2], agungsegara24@gmail.com[3]

Abstrak **— Udara yang bersih merupakan kebutuhan yang tidak hanya perlu dipenuhi manusia, tetapi juga hewan dan tumbuhan. Kualitas udara menjadi penting untuk kehidupan kita agar bisa terhindar dari berbagai penyakit kesehatan. Maka dari itu, diterapkanlah ilmu dari *data mining* dengan metode *k-means clustering* dengan perangkat lunak Orange untuk bisa mengelompokkan kategori dari kualitas udara yang berada di DKI Jakarta yang meliputi kualitas udara baik, sedang, dan tidak sehat. Dengan menggunakan data yang diperoleh dari Indeks Standar Pencemaran Udara (ISPU) dari Dinas Lingkungan Hidup Provinsi DKI Jakarta tahun 2021, didapat hasil bahwa *cluster 0* atau kategori kualitas udara sedang sebanyak 153 item dan *cluster 1* atau udara tidak sehat sebanyak 212 item. Tingkat akurasi dari implementasi *Cluster K-Means* untuk menentukan kualitas udara di DKI Jakarta adalah 0,9622 atau 96,22%. Dengan hasil berikut maka diharapkan pula kepada masyarakat untuk bisa mengurangi pemakaian kendaraan bermotor yang dapat menyebabkan pencemaran udara di DKI Jakarta.**

Kata kunci— ***Data Mining, Clustering,*** **Algoritma *K-Means*, DKI Jakarta, Pencemaran Udara**

I. PENDAHULUAN

Kualitas udara yang bersih merupakan suatu kebutuhan yang tidak hanya dibutuhkan oleh manusia, tetapi juga hewan dan tumbuhan. Pedesaan merupakan salah satu contoh tempat di mana kita dapat memperoleh kualitas udara yang baik karena masih memiliki banyak pepohonan. Selain itu, pedesaan juga memiliki polusi yang berasal dari kendaraan bermotor dan asap pabrik yang lebih sedikit dibanding di daerah perkotaan.

Kesehatan lingkungan adalah kesehatan yang mengatur keseimbangan kesehatan yang dimiliki oleh manusia di sekitar lingkungannya. Masyarakat sepatutnya mewaspadai akibat dari masalah polusi udara yang terjadi di kota-kota besar dan tidak meremehkannya, masih banyak di antara masyarakat yang tetap memaksakan diri untuk keluar rumah dimana mereka mengetahui juga bahwa indikator polusi udara yang tercemar di jalan bahwa udara sangatlah tidak sehat dan berbahaya bagi kesehatan. Untuk itu, masyarakat harus sadar akan akibat dari masalah polusi udara, terutama kita harus tahu kapan saja kondisi polusi udara berbahaya bagi kesehatan manusia, yakni salah satu caranya yaitu dengan memprediksi dengan melihat pola dari polusi udara yang telah terjadi beberapa tahun terakhir sehingga masyarakat dapat lebih waspada dan dapat mencegah akibat dari polusi udara.

II. LANDASAN TEORI

Gas polutan merupakan suatu gas yang menjadi faktor penyebab utama pada polusi udara. Gas ini selalu menimbulkan penyakit yang menyasar makhluk hidup baik tumbuhan hewan, dan manusia.

Berdasarkan peraturan pemerintah RI No.41 Tahun 1999 tentang Pengendalian Pencemaran Udara, bahwasanya yang pencemaran udara adalah masuknya zat, energi atau komponen lain yang sejenis ke dalam udara oleh aktifitas manusia sehingga kadar mutu udara menjadi turun hingga menyebabkan udara tidak sesuai fungsinya lagi.

Data Mining adalah suatu proses pengumpulan informasi dari suatu data yang besar. *Clustering* merupakan proses pengelompokan data menjadi kelompok-kelompok sehingga data di suatu kelompok tersebut memiliki keseragaman jenis. Dengan teknik ini data dapat mengelompokkan kelas atau *cluster*, di mana objek yang memiliki kesamaan untuk dapat dibandingkan dengan objek lainnya.

Algoritma *K-means* ialah algoritma yang dapat mengklasifikasikan jenis dan kelompok data apa berdasarkan titik *centroid* terdekat dengan data informasinya. Tujuan dari algoritma *K-means* ini adalah pengelompokkan data dengan mengupayakan kesamaan atau kemiripan data informasi didalam klaster-klaster. Ukuran kemiripan daya yang digunakan adalah adalah fungsi jarak. Mana jarak terpendek antara data terhadap titik *centroid*, maka itulah yang

dikategorikan data yang mirip.

Penulis telah melakukan olah data dan menggalian data informasi terkait pencemaran udara di DKI Jakarta. Berikut data yang dapat dijadikan ukuran parameter pencemaran udara.

Kategori	Rentang	Carbon Monoksida (CO)	Nitrogen (NO_2)	Ozon (O_3)	Sulfur Dioksida (SO_2)	Partikulat
Baik	0 – 50	Tidak ada efek	Sedikit berbau	Luka pada beberapa spesies tumbuhan akibat kombinasi dengan SO_2 (selama 4 jam)	Luka pada beberapa spesies tumbuhan akibat kombinasi dengan O_3 (selama 4 jam)	Tidak ada efek
Sedang	51 – 100	Perubahan kimia darah tapi tidak terdeteksi	Berbau	Luka pada beberapa spesies tumbuhan	Luka pada beberapa spesies tumbuhan	Terjadi penurunan pada jarak pandang
Tidak Sehat	101 – 199	Peningkatan pada kardiovaskular pada perokok yang sakit jantung	Bau dan kehilangan warna. Peningkatan reaktivitas pembuluh tenggorokan pada penderita asma	Penurunan kemampuan pada atlit yang berlatih keras	Bau, meningkatnya kerusakan tanaman	Jarak pandang turun dan terjadi pengotoran debu dimana-mana
Sangat Tidak Sehat	200 – 299	Meningkatnya kardiovaskular pada orang bukan perokok yang berpenyakit jantung, dan akan tampak beberapa kelemahan yang terlihat secara nyata	Meningkatnya sensitivitas pasien yang berpenyakit asma dan bronhitis	Olah raga ringan mengakibatkan pengaruh pemafasan pada pasien yang berpenyakit paru-paru kronis	Meningkatnya sensitivitas pada pasien berpenyakit asma dan bronhitis	Meningkatnya sensitivitas pada pasien berpenyakit asma dan bronhitis
Berbahaya	300 – lebih	Tingkat yang berbahaya bagi semua populasi yang terpapar				

Gambar 1. Parameter pencemar

Tahap-tahap dari algoritma *K-means clustering* adalah sebagai berikut:

1. Menentukan jumlah *cluster* (k) yang diinginkan sebagai input.
2. Tentukan titik pusat *cluster*/*centroid* secara acak sebanyak.
3. Hitung jarak antara data dengan *centroid*. Pada penelitian ini menggunakan *Euclidean Distance* yaitu metode paling popular untuk mencari jarak terpendek antara data dengan *centroid* dengan rumus yaitu :

$$D(Xi,Yj) = \sqrt{(P1i - Q1j)2}$$

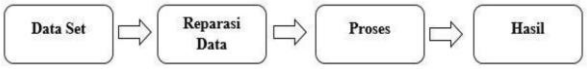
Dimana:
D(X*i*,Y*j*) = Jarak data *i* ke centroid *j*
P1*i* = Variabel - 1 dari data ke *i*
Q1*j* = Variabel - 1 dari data ke *j*

4. Kelompokkan data berdasarkan jarak terpendeknya antara data dengan *centroid* menjadi sebuah kelompok *cluster*.
5. Hitung rata-rata tiap kelompok *cluster* yang terbentuk untuk dijadikan sebagai *centroid* yang baru dan ulangi perhitungan mencari jarak terpendek antara data dan *centroid* apabila *centroid* berubah dan perhitungan akan berhenti apabila *centroid* tidak mengalami perubahan.

Aplikasi Orange merupakan perangkat lunak yang dapat dijadikan suatu alat untuk dapat mengukur pencemaran udara yang bersifat *opensource*. Orange menyediakan pemrograman visual yang berbasis pada komponen-komponen untuk penambangan data, analisis data, pembelajaran mesin dan visualisasi data. Aplikasi Orange sudah dikembangkan sejak tahun 1996 di Ljubljana University dan Jožef Stefan Institute.

III. METODOLOGI PENELITIAN

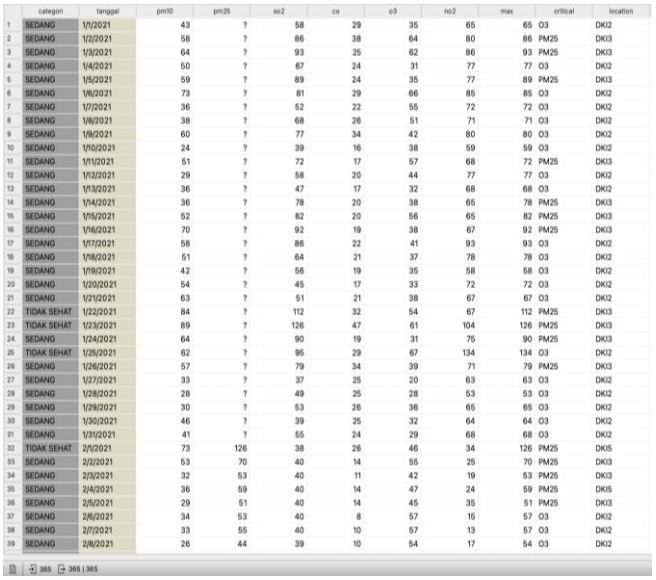
Metodologi penelitian adalah sistematika keseluruhan tahapan yang akan di laksanakan selama penelitian. Langkah-langkah dari metodologi penelitian yaitu sebagai berikut :



Gambar 2. Tahapan Metodologi Penelitian

1. Dataset Indeks Standar Pencemaran Udara (ISPU)
   Dataset yang diambil untuk proses *data mining* diambil dari Dinas Lingkungan Hidup Provinsi DKI Jakarta.

2. Preparasi Data
   Preparasi data dilakukan melalui proses pembersihan data (*Cleaning*). Proses *cleansing data* dilakukan untuk menghilangkan data yang tidak konsisten, atau menghapus atribut yang tidak diperlukan. Gambar 3 merupakan data yang belum melalui proses *cleansing data*.

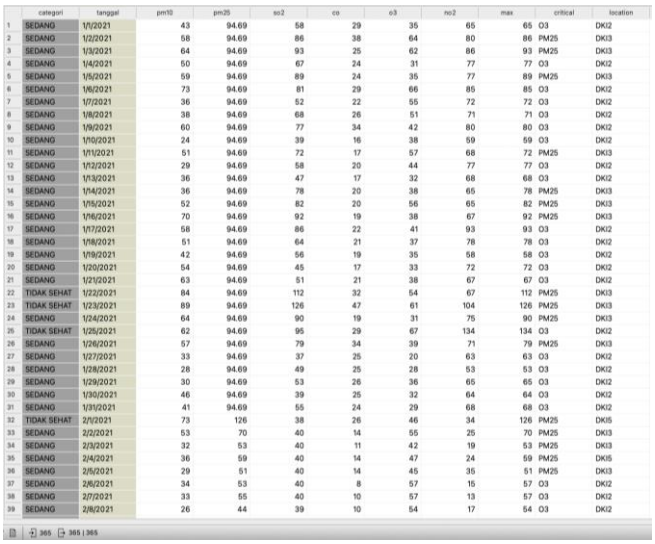


	categori	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	location
1	SEDANG	1/1/2021	43	?	58	29	35	65	65	O3	DKI2
2	SEDANG	1/2/2021	58	?	86	38	64	80	86	PM25	DKI3
3	SEDANG	1/3/2021	64	?	93	25	62	86	93	PM25	DKI3
4	SEDANG	1/4/2021	50	?	67	24	31	77	77	O3	DKI2
5	SEDANG	1/5/2021	59	?	89	24	35	77	89	PM25	DKI3
6	SEDANG	1/6/2021	73	?	81	29	66	85	85	O3	DKI2
7	SEDANG	1/7/2021	36	?	52	22	55	72	72	O3	DKI2
8	SEDANG	1/8/2021	38	?	68	26	51	71	71	O3	DKI2
9	SEDANG	1/9/2021	60	?	77	34	42	80	80	O3	DKI2
10	SEDANG	1/10/2021	24	?	39	16	38	59	59	O3	DKI2
11	SEDANG	1/11/2021	51	?	72	17	57	68	72	PM25	DKI3
12	SEDANG	1/12/2021	29	?	58	20	44	77	77	O3	DKI2
13	SEDANG	1/13/2021	36	?	47	17	32	68	68	O3	DKI2
14	SEDANG	1/14/2021	36	?	78	20	38	65	78	PM25	DKI3
15	SEDANG	1/15/2021	52	?	82	20	56	65	82	PM25	DKI3
16	SEDANG	1/16/2021	70	?	92	19	38	67	92	PM25	DKI3
17	SEDANG	1/17/2021	68	?	86	22	41	93	93	O3	DKI2
18	SEDANG	1/18/2021	61	?	64	21	37	78	78	O3	DKI2
19	SEDANG	1/19/2021	42	?	56	19	35	58	58	O3	DKI2
20	SEDANG	1/20/2021	54	?	45	17	33	72	72	O3	DKI2
21	SEDANG	1/21/2021	63	?	51	21	38	67	67	O3	DKI2
22	TIDAK SEHAT	1/22/2021	84	?	112	32	54	67	112	PM25	DKI3
23	TIDAK SEHAT	1/23/2021	89	?	126	47	61	104	126	PM25	DKI3
24	SEDANG	1/24/2021	64	?	90	19	31	75	90	PM25	DKI3
25	TIDAK SEHAT	1/25/2021	62	?	95	29	67	134	134	O3	DKI2
26	SEDANG	1/26/2021	57	?	79	34	39	71	79	PM25	DKI3
27	SEDANG	1/27/2021	33	?	37	25	20	63	63	O3	DKI2
28	SEDANG	1/28/2021	28	?	49	25	28	53	53	O3	DKI2
29	SEDANG	1/29/2021	30	?	53	26	36	65	65	O3	DKI2
30	SEDANG	1/30/2021	46	?	39	25	32	64	64	O3	DKI2
31	SEDANG	1/31/2021	41	?	55	24	29	68	68	O3	DKI2
32	TIDAK SEHAT	2/1/2021	73	126	38	26	46	34	126	PM25	DKI5
33	SEDANG	2/2/2021	53	70	40	14	55	25	70	PM25	DKI3
34	SEDANG	2/3/2021	32	53	40	11	42	19	53	PM25	DKI3
35	SEDANG	2/4/2021	36	59	40	14	47	24	59	PM25	DKI5
36	SEDANG	2/5/2021	29	51	40	14	45	35	51	PM25	DKI3
37	SEDANG	2/6/2021	34	53	40	8	57	15	57	O3	DKI2
38	SEDANG	2/7/2021	33	55	40	10	57	13	57	O3	DKI2
39	SEDANG	2/8/2021	26	44	39	10	54	17	54	O3	DKI2

365 365 | 365

Gambar 3. Dataset Indeks Standar Pencemaran Udara

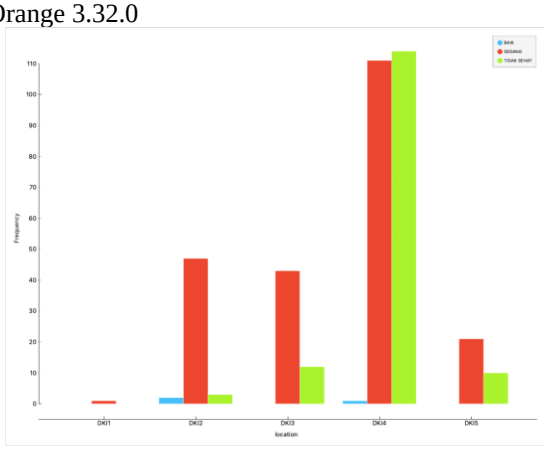
Sedangkan pada gambar 4, merupakan data dari hasil *cleansing data*.



	categori	tanggal	pm10	pm25	so2	co	o3	no2	max	critical	location
1	SEDANG	1/1/2021	43	94.69	58	29	35	65	65	O3	DKI2
2	SEDANG	1/2/2021	58	94.69	86	38	64	80	86	PM25	DKI3
3	SEDANG	1/3/2021	64	94.69	93	25	62	86	93	PM25	DKI3
4	SEDANG	1/4/2021	50	94.69	67	24	31	77	77	O3	DKI2
5	SEDANG	1/5/2021	59	94.69	89	24	35	77	89	PM25	DKI3
6	SEDANG	1/6/2021	73	94.69	81	29	66	85	85	O3	DKI2
7	SEDANG	1/7/2021	36	94.69	52	22	55	72	72	O3	DKI2
8	SEDANG	1/8/2021	38	94.69	68	26	51	71	71	O3	DKI2
9	SEDANG	1/9/2021	60	94.69	77	34	42	80	80	O3	DKI2
10	SEDANG	1/10/2021	24	94.69	39	16	38	59	59	O3	DKI2
11	SEDANG	1/11/2021	51	94.69	72	17	57	68	72	PM25	DKI3
12	SEDANG	1/12/2021	29	94.69	58	20	44	77	77	O3	DKI2
13	SEDANG	1/13/2021	36	94.69	47	17	32	68	68	O3	DKI2
14	SEDANG	1/14/2021	36	94.69	78	20	38	65	78	PM25	DKI3
15	SEDANG	1/15/2021	52	94.69	82	20	56	65	82	PM25	DKI3
16	SEDANG	1/16/2021	70	94.69	92	19	38	67	92	PM25	DKI3
17	SEDANG	1/17/2021	58	94.69	86	22	41	93	93	O3	DKI2
18	SEDANG	1/18/2021	51	94.69	64	21	37	78	78	O3	DKI2
19	SEDANG	1/19/2021	42	94.69	56	19	35	58	58	O3	DKI2
20	SEDANG	1/20/2021	54	94.69	45	17	33	72	72	O3	DKI2
21	SEDANG	1/21/2021	63	94.69	51	21	38	67	67	O3	DKI2
22	TIDAK SEHAT	1/22/2021	84	94.69	112	32	54	67	112	PM25	DKI3
23	TIDAK SEHAT	1/23/2021	89	94.69	126	47	61	104	126	PM25	DKI3
24	SEDANG	1/24/2021	64	94.69	90	19	31	75	90	PM25	DKI3
25	TIDAK SEHAT	1/25/2021	62	94.69	95	29	67	134	134	O3	DKI2
26	SEDANG	1/26/2021	57	94.69	79	34	39	71	79	PM25	DKI3
27	SEDANG	1/27/2021	33	94.69	37	25	20	63	63	O3	DKI2
28	SEDANG	1/28/2021	28	94.69	49	25	28	53	53	O3	DKI2
29	SEDANG	1/29/2021	30	94.69	53	26	36	65	65	O3	DKI2
30	SEDANG	1/30/2021	46	94.69	39	25	32	64	64	O3	DKI2
31	SEDANG	1/31/2021	41	94.69	55	24	29	68	68	O3	DKI2
32	TIDAK SEHAT	2/1/2021	73	126	38	26	46	34	126	PM25	DKI5
33	SEDANG	2/2/2021	53	70	40	14	55	25	70	PM25	DKI3
34	SEDANG	2/3/2021	32	53	40	11	42	19	53	PM25	DKI3
35	SEDANG	2/4/2021	36	59	40	14	47	24	59	PM25	DKI5
36	SEDANG	2/5/2021	29	51	40	14	45	35	51	PM25	DKI3
37	SEDANG	2/6/2021	34	53	40	8	57	15	57	O3	DKI2
38	SEDANG	2/7/2021	33	55	40	10	57	13	57	O3	DKI2
39	SEDANG	2/8/2021	26	44	39	10	54	17	54	O3	DKI2

Gambar 4. Hasil *Cleansing Data*

3. Proses
   Proses *clustering* akan dilakukan pada Orange, yaitu proses pengelompokkan data untuk menentukan mana yang merupakan kategori kualitas udara baik, sedang, dan tidak sehat. Algoritma yang digunakan adalah algoritma *k-means*. Aplikasi yang digunakan adalah Orange 3.32.0

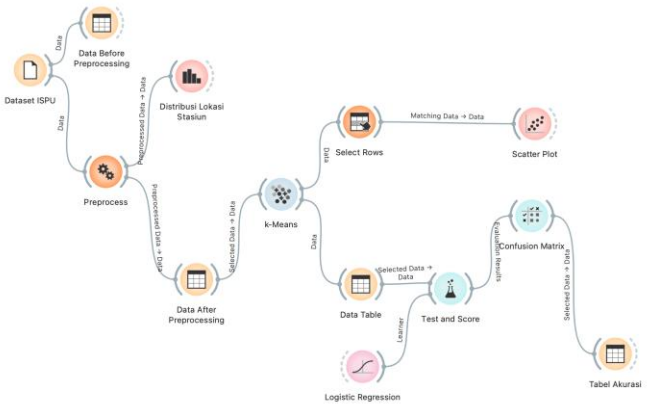


Gambar 5. Kategori Kualitas Udara

4. Hasil
   Dengan data Indeks Standar Pencemaran Udara DKI Jakarta Tahun 2021, akan dikelompokkan kualitas udara yang baik, sedang, dan tidak sehat. Pengelompokkan tersebut berdasarkan atribut berupa parameter seperti PM10, PM25, SO2, CO, O3, dan NO2. Atribut tersebut akan diolah melalui Orange dengan menggunakan algoritma *k-means* yang akan menghasilkan *cluster* dan *centroid*.

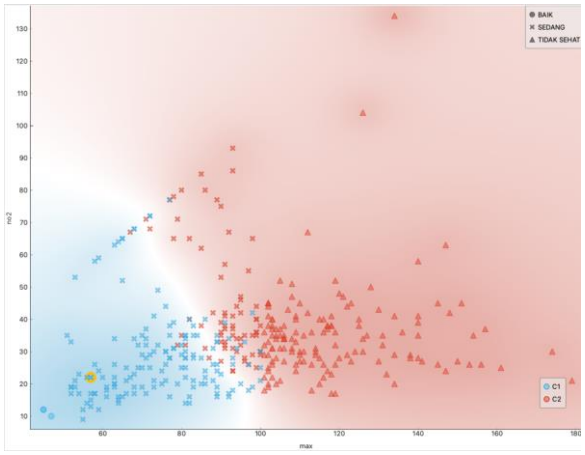
IV. HASIL DAN PEMBAHASAN

A. Pengujian Algoritma *K-Means* pada Orange
   Proses pengujian dilakukan tanpa melakukan perubahan pada aplikasi Orange. Adapun pengujian yang dilakukan yaitu dengan menggunakan algoritma *k-means*. Berikut proses pengujiannya dilakukan sebagai berikut:



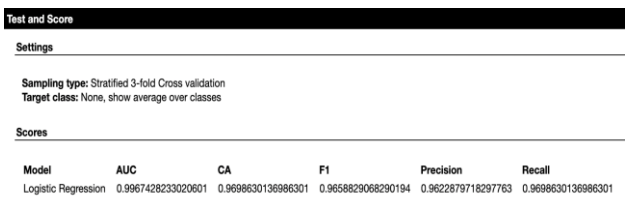
Gambar 6. Pemodelan *Clustering K-Means*

B. Hasil Data *Cluster K-Means*
   Dengan menggunakan pemodelan *k-means clustering* seperti gambar 6 diatas, dengan inisialisasi jumlah *cluster* sebanyak 2 buah, maka didapatkan hasil dengan *cluster* yang terbentuk adalah 2, sesuai dengan pendefinisian nilai k dengan jumlah *cluster* 0 ada 153 item, *cluster* 1 ada 212 item dengan total jumlah data sebanyak 365.



Gambar 7. Hasil *Scatter Plot Cluster K-Means*

Algoritma *cluster k-means* akan diuji dengan *logistic regression*, hasilnya sebagai berikut:



Test and Score

Settings

Sampling type: Stratified 3-fold Cross validation
Target class: None, show average over classes

Scores

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.9967428233020601	0.9698630136986301	0.9658829068290194	0.9622879718297763	0.9698630136986301

Gambar 8. Pengujian Akurasi *Cluster K-Means*

Gambar 9. *Confusion Matrix* dari *Cluster K-Means*

Tingkat akurasi dari implementasi *Cluster K-Means* untuk menentukan kualitas udara di DKI Jakarta adalah 0,9622 atau 96,22%.

V. KESIMPULAN

Dari hasil pengujian yang dilakukan terhadap implementasi *Cluster K-Means* untuk menentukan kualitas udara di DKI Jakarta adalah maka dapat ditarik kesimpulan sebagai berikut:

1. Dalam penelitian ini penulis ingin mengelompokkan data ke dalam kategori berdasarkan kualitas udara di DKI Jakarta, yaitu kualitas baik, kualitas sedang, dan kualitas tidak sehat.
2. Berdasarkan pengujian yang telah dilakukan pada Orange menggunakan algoritma *k-means*, maka didapatkan akurasi berdasarkan pengujian *logistic regression* yaitu 0,9622.
3. Kategori udara berkualitas baik tidak dapat dikelompokkan ke dalam kategorinya sendiri karena berdekatan dengan titik *centroid* kualitas sedang.

REFERENCES

A. Iizuka, S. Shirato, A. Mizukoshi, M. Noguchi, A. Yamasaki dan Y. Yanagisawa, "*A Cluster Analysis of Constant Ambient Air Monitoring Data from the Kanto Region of Japan*," International Journal of Environmental Research and Public Health, pp. 6844-6855, 2014.

B. Warsito, D. Ispriyanti dan H. Widayanti, "*Clustering Data Pencemaran Udara Sektor Industri di Jawa Tengah Dengan Kohonen Neural Network*," Jurnal PRESIPITASI, 2008.

Asroni dan R. Adrian, "*Penerapan Metode K- means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang*," Jurnal Ilmiah Semesta Teknika, pp. 76-82, 2015.

R. Field, "*Indonesia's Dangerous Haze*," 26 October 2015. [Online]. Available: http://www.policyforum.net/indonesias - dangerous-haze/0.

J. Han dan M. Kamber, *"Data Mining Concepts and Techniques Second Edition"*, San Francisco: Morgan Kaufmann, 2006

C. A. Sugianto, "*Analisis Komparasi Algoritma Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan*," Techno.Com, vol. 14, no. 4, pp. 336–342, 2015, doi: 10.33633/tc.v14i4.992.

Sitorus. M, DaCosta. C.A, Larasati. C, *"Penerapan Algoritma K-Means Pada Clustering Vaksinasi Covid-19 Daerah Jawa Timur"*, Jurnal Teknosains Kodepena, Vol. 3, No. 1, pp. 22-30, 2022. e-ISSN 2745-438X. p-ISSN 2745-6129.