

MODEL PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN KUALITAS AIR SUNGAINYA MENGGUNAKAN ALGORITMA K-MEANS++

^{1*}Irwan Budiman, ^{2*}Muliadi, Rendy³

^{1,2,3} Program Studi Ilmu Komputer, FMIPA, Universitas Lambung Mangkurat, Banjarbaru, Indonesia

*Corresponding author: ¹ irwan.budiman@ulm.ac.id, ² muliadi@ulm.ac.id

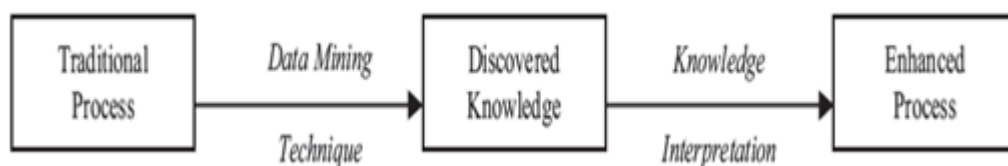
Abstrak. Air sungai mempunyai peranan yang sangat strategis dalam kehidupan manusia dan makhluk hidup lainnya. Air merupakan sumber daya alam yang penting bagi pemenuhan hajat hidup orang banyak sehingga perlu dilindungi agar dapat bermanfaat secara optimal bagi kehidupan. Parameter yang digunakan dalam perhitungan Indeks Kualitas Air yaitu Total Suspended Solid, Dissolved Oxygen, Biochemical Oxygen Demand, Chemical Oxygen Demand, Total Phospat, Fecal Coli dan Total Coli. Data Mining merupakan sebuah proses ekstraksi informasi untuk menemukan pola yang tersembunyi pada tumpukan data dalam database sehingga menjadi pengetahuan yang lebih bermakna. Salah satu fungsi data mining adalah clustering (pengelompokan). Penelitian ini menguji kemampuan data mining khususnya algoritma k-means++ untuk mengelompokkan provinsi di Indonesia berdasarkan kualitas air sungainya.

Kata kunci: kualitas air, data mining, clustering, algoritma k-means++

1. PENDAHULUAN

Air sungai mempunyai peranan yang sangat strategis dalam kehidupan manusia dan makhluk hidup lainnya. Berdasarkan laporan statistik Pusat Data dan Informasi Kementerian Lingkungan Hidup dan Kehutanan tahun 2014, pencemaran air selama tahun 2010 – 2014 menunjukkan kondisi kualitas air sungai umumnya berada pada status tercemar. Pencemaran sungai yang terjadi saat ini semakin beragam dan menunjukkan gejala kerusakan lingkungan yang cukup serius dan dapat menimbulkan konflik pada masyarakat. Kesulitan mendapatkan air bersih dan layak pakai menjadi permasalahan yang sering muncul pada berbagai tempat setiap tahun. Parameter yang digunakan dalam perhitungan Indeks Kualitas Air yaitu Total Suspended Solid, Dissolved Oxygen, Biochemical Oxygen Demand, Chemical Oxygen Demand, Total Phospat, Fecal Coli dan Total Coli.

Data pada bidang lingkungan hidup semakin pesat setiap tahunnya, sehingga menjadi permasalahan tersendiri dalam hal pengelolaannya. Sebuah fakta yang tampak pada institusi lingkungan hidup adalah bahwa pesatnya peningkatan jumlah koleksi data, tidak dibarengi dengan pemanfaatan data tersebut secara maksimal. Proses tradisional jika didukung teknik data mining dapat membantu untuk menemukan pola, struktur dan pengetahuan yang berharga, yang dapat dimanfaatkan oleh instansi terkait sebagai modal untuk peningkatan manajemen khususnya manajemen pengambilan keputusan.



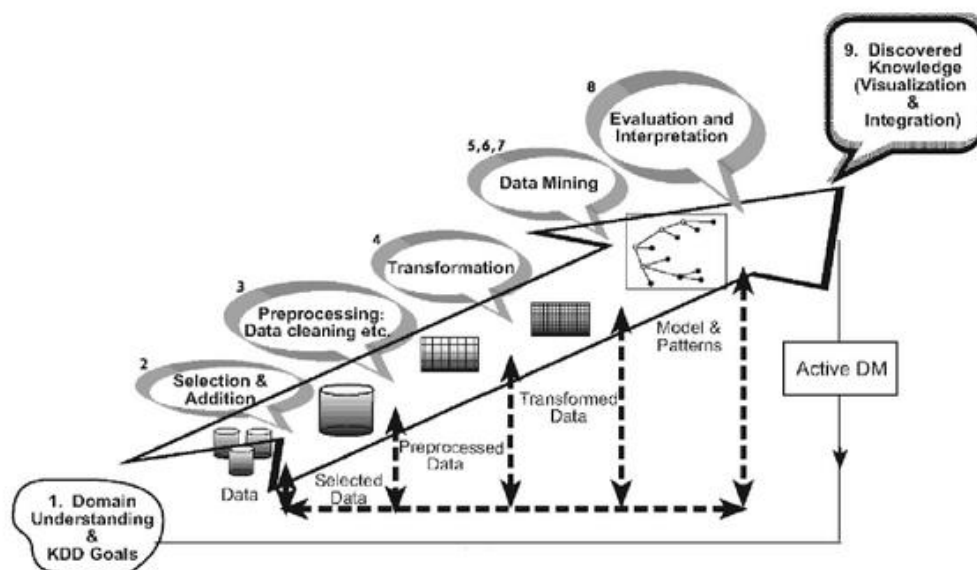
Gambar 1 Fungsi data mining pada penemuan pengetahuan dalam data

Clustering atau pengelompokan menggunakan algoritma K-Means++ merupakan teknik data mining yang berfungsi untuk mengelompokkan data berdasarkan kemiripan data pada suatu kelompok dan meminimalkan kemiripan pada kelompok lain. Sehingga dapat diketahui pola yang tersembunyi dan menemukan pengetahuan tentang tipologi pola kerusakan atau pencemaran air sungai di Indonesia.

2. METODE

Data kualitas air sungai tiap provinsi diproses dengan menerapkan prinsip Knowledge Discovery in Database (KDD) dengan tahapan sebagai berikut:

1. Pengumpulan Data. Data yang digunakan adalah data kualitas air sungai seluruh provinsi di Indonesia berdasarkan statistik Kementerian Lingkungan Hidup dan Kehutanan.
2. Pemilihan Data. Pemilihan tabel database yang diperoleh dengan pemilihan dan penyeleksian atribut yang relevan yang nantinya akan digunakan dalam tahap pemrosesan data mining.
3. Preprocessing Data. Dari data yang telah dipilih, pembersihan akan dilakukan guna menangani nilai yang hilang, induksi noise ataupun outlier. Selain itu, penggunaan query untuk mendapatkan data yang diharapkan juga dilakukan dalam tahapan ini.
4. Transformasi Data. Dalam tahapan ini, data kualitas air sungai akan dilakukan transformasi menjadi kesatuan data dalam tabel agar siap untuk diproses dalam proses data mining menggunakan algoritma K-Means++. Artinya data harus berupa data kuantitatif agar dapat diproses menggunakan algoritma K-Means++.
5. Data Mining. Data mining merupakan tahapan proses dalam menemukan suatu pola atau pengetahuan dalam data yang berjumlah besar. Dalam penelitian ini akan digunakan teknik clustering menggunakan algoritma K-Means++.
6. Pattern Evaluation. Pattern evaluation merupakan tahapan hasil dari teknik data mining berupa evaluasi dasar terhadap kinerja model clustering yang telah dihasilkan. Identifikasi pola-pola khusus yang dapat memberikan representasi pengetahuan berdasarkan pengukuran tertentu. Pada penelitian ini, untuk metode K-Means++ akan dihasilkan berupa cluster atau pengelompokan produk dan akan dievaluasi menggunakan nilai BCV/WCV.
7. Knowledge Presentation. Tahap memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Tahapan ini dilakukan visualisasi dan presentasi hasil pengetahuan yang diperoleh. Visualisasi dapat disajikan dalam bentuk tabel dan grafik dari hasil yang didapat.



Gambar 2 Prosedur Penelitian Berdasarkan Proses KDD

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini yaitu data indeks kualitas air sungai yang diperoleh dari buku Statistik Lingkungan Hidup Indonesia Yang diterbitkan dari tahun 2014-2018. Data kualitas air sungai mempunyai jumlah data sebanyak 33 jumlah data sesuai dengan jumlah provinsi yang ada di Indonesia.

3.2. Pemilihan Data, Preprocessing Data dan Transformasi Data

Dari data yang cukup banyak tersebut dilakukan pemilihan data, kemudian preprocessing data dan transformasi data maka dihasilkan sejumlah atribut pada data indeks kualitas air sungai ini ada 7 atribut data yaitu Total suspended solid (mg/l), Dissolved Oxygen (mg/l), Biochemical Oxygen Demand (mg/l), Chemical Oxygen Demand (mg/l), Fecal Coli (jml/100 ml), Total Coli (jml/100 ml) dan Total Fosfat (mg/l).

Tabel 1 Data Indeks Kualitas Air Sungai di Seluruh Ibukota Provinsi di Indonesia

No.	Provinsi	Sungai	TSS	DO	BOD	COD	Fecal Coli	Total Coli	Total Fosfat
1	Banda Aceh	Krueng Aceh	33,5	5,62	1,48	16	785	1750	0,23
2	Medan	Deli	71	6,39	13,22	24,2	5	100	0,67
3	Padang	Batang Hari	13,83	5,74	2,42	11,68	3211,11	10910	0,06
4	Pekanbaru	Siak	28,58	3,36	4,68	39,64	428,88	1950,91	0,07
5	Jambi	Batang Hari	41	6,04	2,48	26,17	633,33	1995	0,1
6	Palembang	Musi	28,03	5,33	1,58	9,36	55,56	1677,78	0,13
7	Bengkulu	Sungai Bengkulu	575,91	5,17	1,43	12,04	989,8	2150	0,28
8	Bandar Lampung	Way Sekampung Hulu	6,5	6,31	16,75	28,35	75	247,75	0,1
9	Pangkal Pinang	Baturusa	10,58	4,86	2,66	8,99	20,69	167,25	0,06
10	Tanjung Pinang	-	5,8	6,47	4	0,3	38,73	126,33	0,13
11	Jakarta	Krukut	63,17	2,87	17,26	110,63	122180	122180	0,57
12	Bandung	Ciliwung	13,89	3,37	5,3	36,68	182620	559443	0,12
13	Semarang	Garang	45,33	6,8	3,07	7,18	59500	338500	0,08
14	Yogyakarta	Code	15,75	6,88	4,64	11,49	12375	40333,3	0,26
15	Surabaya	Brantas	501,56	6,52	5,81	13,14	3900	7500	0,26
16	Serang	Cibanten	22,15	2,61	5,76	57	3093,11	4216,9	0,04
17	Denpasar	Tukad Badung	60,67	5,57	6,19	19,76	574650	1261167	0,25
18	Mataram	Jangkak	15,33	6,7	8,11	24,35	160000	135933	0,18
19	Kupang	Noelmina	26,4	7,11	8,9	14,24	459,54	12852,9	0,46
20	Pontianak	Kapuas	22	6,36	7,37	37,92	494,63	130	0,04
21	Palangka Raya	Kahayan	229,27	6,19	4,99	33,37	857	861	0,19
22	Banjarmasin	Barito	40,99	4,76	5,9	13,57	1732,61	13103,7	0,18
23	Samarinda	Mahakam	27,06	5,51	1,24	3,57	3877,16	14258,5	0,03
24	Manado	Tondano	2,07	0,25	0,4	0,67	19,467	5,38	0,23
25	Palu	Lariang	36	5	6,5	26,33	124150	10016	0,08
26	Makassar	Sa'dan	675,57	6,22	2,97	24,67	562,57	1121,14	0,11
27	Kendari	Wanggu	16,92	6,34	3,98	12,2	14,2	129	0,01
28	Gorontalo	Bone	34,25	7,64	2	10	17544,2	24200	0,15
29	Mamuju	Lariang	59,67	7,72	4,14	13,64	4644,44	5897,78	0,18
30	Ambon	Wae Batu Gajah	0,33	7,6	2,53	4,68	951	108	0,19
31	Ternate	Tabobo	1	7,96	2	10	2545	20590	0,21
32	Manokwari	Maruni	63,4	8,1	2,02	7,08	994	2006	0,09
33	Jayapura	Kampwolker	51	6,03	1,93	13,89	17,2	869,87	0,06

3.3. Data Mining dan Pattern Evaluation

Proses data mining menggunakan algoritma k-means ++ menghasilkan clustering atau pengelompokan data menghasilkan 3 kelompok provinsi berdasarkan kualitas air sungainya. Dengan menggunakan pusat cluster awal seperti tabel berikut:

Tabel 2. Pusat cluster awal

Indeks	C01	C02	C03	C04	C05	C06	C07
15	501,56	6,52	5,81	13,14	3900	7500	0,26
17	60,67	5,57	6,19	19,76	574650	1261167	0,25
2	71	6,39	13,22	24,2	5	100	0,67

Dan dengan melalui 6 kali iterasi proses algoritma k-means++ menghasilkan pusat cluster akhir sebagai berikut:

Tabel 3. Pusat cluster akhir

No	C01	C02	C03	C04	C05	C06	C07
1	29,61	5,085	4,185	21,93	121060	448971,4	0,1
2	60,67	5,57	6,19	19,76	574650	1261167	0,25
3	90,62067	5,788667	4,971667	20,639	15555,14	14579,59	0,178333

3.4. Knowledge Presentation

Berdasarkan hasil algoritma k-means++ provinsi di Indonesia terbagi menjadi 3 kelompok dimana setiap kelompok memiliki kemiripan satu sama lain dalam hal indeks kualitas air sungainya. Adapun sebarannya adalah sebagai berikut:

Tabel 4. Sebaran kelompok provinsi

No.	Provinsi	Sungai	Cluster
1	Banda Aceh	Krueng Aceh	3
2	Medan	Deli	3
3	Padang	Batang Hari	3
4	Pekanbaru	Siak	3
5	Jambi	Batang Hari	3
6	Palembang	Musi	3
7	Bengkulu	Sungai Bengkulu	3
8	Bandar Lampung	Way Sekampung Hulu	3
9	Pangkal Pinang	Baturusa	3
10	Tanjung Pinang	-	3
11	Jakarta	Krukut	3
12	Bandung	Ciliwung	1
13	Semarang	Garang	1
14	Yogyakarta	Code	3
15	Surabaya	Brantas	3

16	Serang	Cibanten	3
17	Denpasar	Tukad Badung	2
18	Mataram	Jangkok	3
19	Kupang	Noelmina	3
20	Pontianak	Kapuas	3
21	Palangka Raya	Kahayan	3
22	Banjarmasin	Barito	3
23	Samarinda	Mahakam	3
24	Manado	Tondano	3
25	Palu	Lariang	3
26	Makassar	Sa'dan	3
27	Kendari	Wanggu	3
28	Gorontalo	Bone	3
29	Mamuju	Lariang	3
30	Ambon	Wae Batu Gajah	3
31	Ternate	Tabobo	3
32	Manokwari	Maruni	3
33	Jayapura	Kampwolker	3

4. SIMPULAN

Kesimpulan dari penelitian ini, algoritma k-means ++ mampu menghasilkan pengelompokan 3 cluster provinsi di Indonesia berdasarkan kualitas air sungainya dengan 2 provinsi pada cluster pertama yaitu index rendah, 1 provinsi pada cluster kedua yaitu index sedang dan 30 provinsi pada cluster ketiga yaitu index tinggi.

5. UCAPAN TERIMA KASIH

Ucapan terima kasih yang sebesar-besarnya kepada pihak Rektorat yang telah membantu memberikan dana hibah penelitian pada program Dosen Wajib Meneliti tahun 2020. Dan tak lupa ucapan terima kasih kepada Kepala Laboratorium Data Science Program Studi Ilmu Komputer FMIPA ULM yang membantu pelaksanaan penelitian data mining ini.

6. DAFTAR PUSTAKA

- Agusta, Yudi. (2007). *K-Means Penerapan, Permasalahan dan Metode Terkait*. Jurnal Sistem dan Informatika, 3 : 47-60.
- Agrawal, A. & Gupta, H., (2013). *Global K-Means (GKM) Clustering Algorithm: A Survey*. International Journal of Computer Applications, LX(2) : 20-24
- Asroni & R. Adrian. (2015). *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. 18(1):76-82.
- Celebi, M.E., Kingravi, H.A. & Vela, P.A., (2013). *Expert Systems with Applications A comparative study of efficient initialization methods for the K-Means clustering algorithm*. Expert Systems With Applications. 40(1) : 200-210
- Debatty, Thibault. (2014). *Determining the k in K-Means with MapReduce*. Proceedings of the EDBT/ICDT 2014 Joint Conference (ISSN 1613-0073). 19-28.
- Ediyanto, Muhlasah N.M. & Neva S. (2013). *Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis*. Buletin Ilmiah Mat. Stat. dan Terapannya. 02(2) : 133-136.

- Hartigan, J. A., & Hartigan, J. A. (1975). *Clustering algorithms* (Vol. 209). New York: Wiley.
- Madhulatha, T.S. (2012). *An Overview On Clustering Methods*. IOSR Journal of Engineering, 2(4) : 719-725.
- Narwati, (2010). Pengelompokan Mahasiswa Menggunakan Algoritma K-Means. Fakultas Teknologi dan Informasi
- Rencher, A. C. (2002). Principal component analysis. *Methods of Multivariate Analysis, Second Edition*, 380-407.
- Santosa, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu, 978(979), 756.
- Teknomo, K. (2007). K-Means clustering tutorials. *Medicine*, 100(4), 3.