



## Data Mining Klasterisasi dengan Algoritme K-Means untuk Pengelompokan Provinsi Berdasarkan Konsumsi Bahan Bakar Minyak Nasional

Indah Rizky Mahartika<sup>1</sup>, Arief Wibowo<sup>2</sup>

<sup>1,2</sup> Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur

<sup>1</sup>[indahriz@gmail.com](mailto:indahriz@gmail.com), <sup>2</sup>[arief.wibowo@budiluhur.ac.id](mailto:arief.wibowo@budiluhur.ac.id)

### Abstract

Petroleum is one of the natural resources that play an important role in human life, mainly used as the fuel needed by all levels of society. The distribution of fuel oil (BBM) in Indonesia is carried out by the Downstream Oil and Gas Regulatory Agency (BPH Migas). With the availability of data on fuel consumption in each province, it can be seen that the pattern of fuel consumption in Indonesia is beneficial for regulators in the management of fuel distribution. To find out the pattern of national fuel consumption, we need a model of grouping regions in Indonesia based on the level of fuel consumption in each province. This study analyzes data on national fuel consumption throughout Indonesia using the Data Mining Clustering technique, and the Euclidean Distance measurement method. The final results of this study indicate that the K-Means algorithm can group provinces based on national fuel consumption levels into three clusters with their respective specifications. Modeling results were evaluated using the Davies Bouldin Index (DBI) instrument, with a value of 0.32. The results of testing using DBI approaching 0 indicate that the clusters formed are relatively very good and ideal.

Kata kunci: *Data Mining, Clasterisation, K-Means, Davies Bouldin Index, BBM*

### Abstrak

Minyak Bumi merupakan salah satu sumber daya alam yang berperan penting dalam kehidupan manusia, terutama digunakan sebagai bahan bakar yang dibutuhkan oleh seluruh lapisan masyarakat. Penyaluran Bahan Bakar Minyak (BBM) di Indonesia dilakukan oleh Badan Pengantar Hilir Minyak dan Gas (BPH Migas). Dengan tersedianya data mengenai konsumsi BBM di tiap provinsi, maka dapat diketahui pola konsumsi BBM di Indonesia yang bermanfaat bagi regulator dalam tata kelola distribusi bahan bakar minyak. Untuk mengetahui pola konsumsi BBM secara nasional maka dibutuhkan suatu model pengelompokan wilayah yang ada di Indonesia berdasarkan tingkat konsumsi BBM di setiap provinsi. Penelitian ini menganalisis data konsumsi BBM nasional di seluruh Indonesia menggunakan teknik Data Mining Klasterisasi, dan metode pengukuran jarak *Euclidean Distance*. Hasil akhir dari penelitian ini menunjukkan bahwa algoritme K-Means mampu mengelompokkan provinsi berdasarkan tingkat konsumsi BBM nasional menjadi tiga kluster dengan spesifikasi masing-masing. Hasil pemodelan dievaluasi menggunakan instrumen *Davies Bouldin Index (DBI)*, dengan nilai perolehan sebesar 0,32. Hasil pengujian menggunakan DBI yang mendekati 0 menunjukkan bahwa kluster yang terbentuk relatif sangat baik dan ideal.

Kata kunci: *Data Mining, Klasterisasi, K-Means, Davies Bouldin Index, BBM*

### 1. Pendahuluan

Bahan Bakar Minyak (BBM) adalah sumber daya yang berbentuk cairan dan digunakan sebagai sumber energi untuk kendaraan bermotor. Adapun jenis BBM di Indonesia antara lain adalah minyak tanah, petralite, pertamax, solar, premium, avtur, dan sebagainya. Selain itu, ada pula 3 tipe BBM di Indonesia yang diatur oleh pemerintah. Yaitu yang disubsidi oleh pemerintah, non-subsidi, dan penugasan.

Indonesia merupakan negara kepulauan yang luas. Terbang dari Sabang sampai Merauke, dari Niangas samapi Pulau Rote. Hal tersebut tentu menjadi kendala terhadap pendistribusian dan penyediaan bahan bakar minyak. Selain itu, tingkat konsumsi terhadap bahan bakar minyak di setiap wilayah berbeda.

Maka dibutuhkan suatu sistem yang dapat memberi solusi pada dua masalah tersebut. Sistem yang diharapkan mampu mengoptimalkan pemanfaatan data penyaluran BBM yang kemudian dapat juga membantu memperlancar dan mengoptimalkan alokasi kuota penyaluran BBM ke seluruh provinsi di Indonesia. Sehingga digunakan algoritme K-Means untuk mengklasterisasi provinsi berdasarkan tingkat konsumsi BBM.

### 2. Metode Penelitian

#### 2.1. Data Mining

Data *explosion* merupakan latar belakang munculnya data *mining*. Jumlah data yang tersimpan dalam basis data akan semakin membesar dan hanya disimpan begitu saja sebagai laporan tanpa pemanfaatan lebih lanjut.

Data *mining* berusaha memanfaatkan data tersebut dengan melakukan suatu proses yang dapat menghasilkan pola-pola tertentu yang sifatnya tersembunyi. Kemudian pola-pola tersembunyi tersebut dapat menjadi informasi atau pengetahuan yang bermanfaat.

Data *mining* merupakan serangkaian proses dalam pencarian pola, hubungan, penggalian nilai tambah dari data dan informasi yang berukuran besar berupa pengetahuan dengan tujuan menemukan hubungan dan menyederhanakan data agar diperoleh informasi yang dapat dipahami dan bermanfaat dengan bantuan ilmu statistik dan matematika [1].

## 2.2. Klasterisasi

Menurut Nango, D. N. (2012) *Clustering* atau Klasterisasi adalah suatu alat bantu pada data *mining* yang bertujuan untuk mengelompokkan objek-objek ke dalam beberapa klaster. Klaster adalah sekelompok atau sekumpulan objek-objek data yang memiliki kemiripan karakteristik satu sama lain dalam klaster yang sama dan berbeda karakteristik terhadap objek-objek yang berbeda klaster [2].

Cara kerja teknik ini ialah mengelompokkan sekumpulan data ke dalam kelas-kelas atau kluster-kluster, yang mana objek-objek yang ada pada kelas tersebut memiliki similaritas yang tinggi jika dibandingkan dengan objek lain yang ada dalam kelas tersebut, namun memiliki similaritas yang rendah jika dibandingkan dengan objek yang ada di kelas/kluster lain [3].

## 2.3. Algoritme K-Means

Metode K-Means merupakan metode yang termasuk dalam algoritme *clustering* berbasis jarak yang membagi data ke dalam sejumlah klaster dan algoritme ini hanya bekerja pada atribut numerik [4].

Metode dasar analisis algoritme K-Means *Clustering* adalah sebagai berikut [5]:

- Tentukan jumlah klaster ( $k$ ), tetapkan pusat klaster secara acak.
- Hitung jarak setiap data ke pusat klaster.
- Kelompokan data ke dalam klaster dengan jarak yang paling pendek.
- Hitung pusat klaster baru.
- Ulangi langkah 2 (dua) sampai 4 (empat) hingga sudah tidak ada lagi data yang berpindah ke klaster yang lain.

Proses *clustering* dimulai dengan mengidentifikasi data yang diklasterisasi, dapat digunakan rumus formula *Euclidean Distance* seperti yang terlihat pada rumus persamaan (1), berikut [6]:

$$d_{ij} = \sqrt{\sum_{k=1}^m X_{ij} - C_{jk}^2} \quad (1)$$

Dimana  $D_{ij}$  merupakan jarak objek antar nilai data dan nilai pusat *cluster*,  $m$  adalah jumlah dimensi data,  $X_{ij}$  merupakan nilai data dari dimensi ke- $k$  dan  $C_{jk}$  adalah nilai pusat *cluster* dari dimensi ke- $k$ .

Untuk menghitung centroid baru, dapat menggunakan rumus persamaan (2), sebagai berikut:

$$C = \frac{\sum m}{n} \quad (2)$$

Dimana  $C$  merupakan centroid data,  $m$  adalah anggota data yang termasuk ke dalam centroid tertentu dan  $n$  adalah jumlah data yang menjadi anggota *centroid* tertentu.

## 2.4. Knowledge Discovery in Database

Data *mining*, sering juga disebut sebagai *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [7]. Secara garis besar KDD dapat dijelaskan sebagai berikut [8]:

### Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses Data mining disimpan dalam suatu berkas, terpisah dari basis data operasional.

### Pre-processing / Cleaning

Sebelum proses data *mining* dapat dilaksanakan, perlu dilakukan proses pembersihan pada data yang menjadi fokus KDD. Proses pembersihan mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

### Transformation

*Coding* adalah transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data *mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

### Data Mining

Data *mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

### Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses data *mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.5. Davies Bouldin Index

Indeks validitas Davies Bouldin (DB) menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah jumlah nilai *compactness* yang dibagi dengan jarak antara kedua titik pusat kluster sebagai *separation* [9].

Dalam proses evaluasi dari model yang dihasilkan, digunakan Davies Bouldin Index. DBI digunakan untuk memaksimalkan jarak *inter-cluster* dan meminimalkan jarak *intra-cluster* yang dapat dihitung dengan persamaan 3 berikut:

$$S_i = \frac{1}{|c_i|} \sum_{x \in c_i} \{ |x - z_i| \} \tag{3}$$

Dimana  $c_i$  sebagai banyaknya titik yang masuk ke dalam kluster  $i$ ,  $x$  adalah data, dan  $z_i$  centroid dari kluster  $i$ . Sedangkan jarak antara kluster didefinisikan pada Persamaan 4 berikut:

$$d_{ij} = |z_i - z_j| \tag{4}$$

Dimana  $z_i$  centroid dari kluster  $i$  dan  $z_j$  centroid dari kluster  $j$ . Perhitungan jarak  $d_{ij}$  dapat menggunakan euclidean. Selanjutnya akan mendefinisikan  $R_{i,qt}$  untuk kluster  $c_i$  pada Persamaan 5 berikut:

$$R_{i,qt} = \max_{j, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \tag{5}$$

Selanjutnya *Davies Bouldin Index* didefinisikan pada Persamaan 6 berikut:

$$DBI = \frac{1}{k} \sum_{i=1}^k R_{i,qt} \tag{6}$$

Dari persamaan tersebut,  $k$  adalah jumlah cluster yang digunakan. Semakin kecil nilai DBI yang diperoleh (non-negatif  $\geq 0$ ), maka semakin baik cluster yang diperoleh dari pengelompokan K-means yang digunakan [10].

3. Hasil dan Pembahasan

Langkah-langkah untuk melakukan data *mining* mengikuti aturan KDD adalah sebagai berikut:

Data Selection

Data penyaluran BBM yang diolah merupakan data tahun 2018 dan memiliki atribut-atribut sebagai berikut:

- 1) Tahun
- 2) Bulan
- 3) Tipe

Terdapat 3 tipe BBM, antara lain adalah Jenis BBM Umum (JBU). Yaitu Petralite, Pertamina, Pertadex, Dex, Dexlite. Jenis BBM Tertentu (JBT). Yaitu Solar Subsidi dan Minyak Tanah. Jenis BBM Khusus Penugasan (JBKP).

- 4) Nama Provinsi
- 5) Kabupaten

- 6) Sektor
- 7) Generik
- 8) Volume

Dari beberapa atribut yang telah disebutkan sebelumnya, dipilih hanya tiga atribut. Yaitu:

- 1) Nama Provinsi
- 2) Tipe (hanya tipe JBU)
- 3) Volume

Pre-processing / Cleaning

Langkah ini dilakukan untuk membersihkan data duplikasi dan data yang tidak konsisten.

Transformation

Transformasi dilakukan untuk menyesuaikan data agar dapat diproses. Transformasi yang dilakukan adalah atribut Tipe dipilih hanya JBU dan kemudian ditransformasi menjadi jumlah total JBU. Begitu juga dengan volume, yang digunakan adalah jumlah totalnya.

Data Mining

Data *mining* dilakukan dengan tujuan untuk mencari informasi atau pola untuk *clustering* menggunakan algoritme K-Means.

Pada proses data mining menggunakan algoritme K-Means dilakukan beberapa tahapan, yaitu:

- a. Tentukan jumlah kluster ( $k$ ), tetapkan pusat kluster (centroid) secara acak.  
Dibentuk 3 kluster ( $k=3$ ) dengan titik pusat awal sebagai berikut:

Tabel 1 Titik Pusat Kluster (Centroid) Awal

Centroid	Provinsi	JBU	Volume
C1	Bengkulu	293	15012294431
C2	Riau	365	15072018926
C3	Kalimantan Tengah	602	14328426055

- b. Hitung jarak setiap data ke pusat kluster. Perhitungan menggunakan rumus persamaan (1).

Jarak data 1 (Aceh) ke C1 =

$$\sqrt{(248 - 293)^2 + (14032552002 - 15012294431)^2} = 979742429$$

Jarak data 1 (Aceh) ke C2 =

$$\sqrt{(248 - 365)^2 + (14032552002 - 15072018926)^2} = 1039466923$$

Jarak data 1 (Aceh) ke C3 =

$$\sqrt{(248 - 602)^2 + (14032552002 - 14328426055)^2} = 295874052$$

- c. Kelompokkan data ke dalam kluster dengan jarak yang paling pendek.

Seperti contoh data 1 di atas, provinsi Aceh masuk ke dalam kluster 3 karena nilai jarak ke pusat kluster 3 merupakan jarak terkecil diantara jarak terhadap titik pusat kluster yang lainnya.

- d. Hitung pusat kluster (centroid) baru.  
Untuk menghitung pusat kluster baru digunakan rumus persamaan (2).

C1 baru:

$$JBU = 542$$

$$Volume = 14893592418$$

C2 baru:

$$JBU = 386$$

$$Volume = 15077664960$$

C3 baru:

$$JBU = 272$$

$$Volume = 14075483792$$

- e. Ulangi langkah 2 (dua) sampai 4 (empat) hingga sudah tidak ada lagi data yang berpindah ke kluster yang lain. Dapat juga dilihat dari nilai titik pusat yang diperoleh dari setiap iterasinya.

f.

Tabel 2 Centroid Iterasi 1

Centroid	JBU	Volume
C1	542	14.893.592.418
C2	386	15.077.664.960
C3	272	14.075.483.792

Tabel 3 Centroid Iterasi 2

Centroid	JBU	Volume
C1	640	14.806.817.298
C2	357	15.051.033.941
C3	257	14.046.659.628

Tabel 4 Centroid Iterasi 3

Centroid	JBU	Volume
C1	700	14.747.476.140
C2	387	15.029.730.887
C3	257	14.046.659.628

Tabel 5 Centroid Iterasi 4

Centroid	JBU	Volume
C1	700	14.747.476.140
C2	387	15.029.730.887
C3	257	14.046.659.628

Kemudian iterasi dihentikan dan hasil klusterisasi didapat dari iterasi terakhir.

Tabel 6 Hasil Klusterisasi

Kluster	Rata-Rata Konsumsi		Jumlah Provinsi	Anggota Cluster
	JBU	Volume		
1	700	14.747.467.140	5	Banten, Daerah Istimewa Yogyakarta, Jawa Barat, Jawa Tengah, Jawa Timur
2	387	15.029.730.887	9	Bengkulu, DKI Jakarta, Jambi, Kepulauan Bangka Belitung, Kepulauan Riau,

Kluster	Rata-Rata Konsumsi		Jumlah Provinsi	Anggota Cluster
	JBU	Volume		
				Riau, Sumatera Barat, Sumatera Selatan, Sumatra Utara
3	257	14.046.659.628	20	Aceh, Bali, Gorontalo, Kalimantan Barat, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Timur, Kalimantan Utara, Lampung, Maluku, Maluku Utara, Nusa Tenggara Barat, Nusa Tenggara Timur, Papua, Papua Barat, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Tenggara, Sulawesi Utara

Dari data tabel di atas yang dihitung dari centroid hasil iterasi terakhir dibagi banyaknya anggota *cluster* maka diketahui bahwa *kluster 1* merupakan kelompok provinsi dengan pesanan JBU rendah dan volume BBM-nya juga rendah, sedangkan *kluster 2* merupakan sekumpulan provinsi dengan pesanan JBU tinggi namun volume BBM-nya sedang, dan *kluster 3* merupakan kelompok provinsi dengan pesanan JBU sedang tapi volume BBM-nya tinggi.

*Interpretation / Evaluation*

Langkah terakhir dari KDD ini dilakukan untuk menjelaskan makna dari setiap kluster yang terbentuk. Langkah ini dilakukan dengan cara membagi centroid dari iterasi terakhir dengan jumlah anggota setiap klasternya. Sehingga dapat disimpulkan deskripsi dari setiap kluster.

Selain itu pada langkah ini juga melakukan evaluasi kluster dengan metode *Davies Bouldin Index* untuk menghitung rata-rata nilai setiap titik pada himpunan data.

Untuk menghitung nilai DBI, digunakan persamaan 3, 4, 5 dan 6 secara berurutan.

$$S_1 = 2658636773$$

$$S_2 = 3504082556$$

$$S_3 = 8847641329$$

$$M_{1,2,3} = 15730548228$$

$$R_{1,2,3} = \frac{(2658636773 + 3504082556 + 8847641329)}{15730548228}$$

$$= 0,95$$

$$DB = \frac{1}{3} \times 0,95 = 0,32$$

Evaluasi kluster menggunakan perhitungan *Davies Bouldin Index* (DBI) diperoleh hasil 0,32. Hitungan selengkapnya terlampir. Jika nilai DBI semakin kecil atau semakin mendekati 0, maka hasil kluster yang diperoleh semakin bagus dan menunjukkan bahwa hasil kluster yang diperoleh adalah relatif sangat baik.

### 3.1. Rancangan Sistem

Sistem dirancang berbasis *web* dan menggunakan bahasa pemrograman PHP. Berikut tampilan rancangan layar untuk sistem klusterisasi provinsi berdasarkan tingkat BBM dilihat dari jumlah JBU dan volumenya

Provinsi	Konsumsi JBU (by order)	Volume (liter)	Cluster 1	Cluster 2	Cluster 3
Aceh	248	14032552002	✓		
Bali	132	13980523396	✓		
Banten	563	14651367073	✓		
Bengkulu	293	15012294431			✓
Daerah Istimewa Yogyakarta	107	14757464513		✓	
DKI Jakarta	735	14940086876			✓
Gorontalo	46	13936956539	✓		
Jambi	225	15045747860			✓

Gambar 1 Hasil Klusterisasi Tabel Iterasi

Ketika tombol Proses ditekan maka setelah itu akan tampil halaman Hasil Klusterisasi yang menampilkan beberapa tabel iterasi dan tabel terakhir menunjukkan hasil akhir dari proses klusterisasi.

Sulawesi Tengah	160	13984487794	✓		
Sulawesi Tenggara	223	13944251037	✓		
Sulawesi Utara	203	13995119680	✓		
Sumbawa Barat	251	15078225979			✓
Sulawesi Selatan	403	15029449669			✓
Sulawesi Utara	702	15114766860			✓

Centroid JBU dan Volume dari Hasil Terakhir:  
 Cluster 1 -> 257 14.046.658.838  
 Cluster 2 -> 700 14.747.478.140  
 Cluster 3 -> 367 15.030.730.867  
 Nilai Evaluasi DBI : 0,32

Gambar 2 Centroid Terakhir dan Nilai DBI

### 4. Kesimpulan

Kesimpulan yang didapat dari penelitian ini antara lain yaitu pengelompokan provinsi di Indonesia berdasarkan tingkat konsumsi BBM dapat diselesaikan dengan Algoritme K-Means. Algoritme K-Means dapat melakukan pengelompokan data dalam jumlah yang banyak akan tetapi belum cukup efisien untuk

mengelompokkan secara tepat karena penentuan centroid (titik pusat) pada tahap awal algoritme K-Means sangat mempengaruhi hasil kluster seperti pada hasil pengujian yang dilakukan dengan centroid yang berbeda menghasilkan hasil kluster yang berbeda juga. Dalam penelitian ini, iterasi dilakukan sebanyak empat kali.

Pengujian kluster dilakukan menggunakan Davies Bouldin Index (DBI). Nilai DBI yang diperoleh adalah 0,32 sehingga kluster yang terbentuk sudah relatif baik.

### Daftar Rujukan

- G. Abdurrahman, "Clustering Data Ujian Tengah Semester ( UTS ) Data Mining," *J. Sist. Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 71–79, 2016.
- N. Rohmawati, S. Defiyanti, and M. Jajuli, "Implementasi Algoritma K-Means Dalam Pengklusteran Mahasiswa Pelamar Beasiswa," *J. Ilm. Teknol. Inf. Terap.*, vol. I, no. 2, pp. 62–68, 2015.
- O. Riveranda *et al.*, "K-Means Analysis Klusterisasi Kasus HIV / AIDS di Indonesia," no. September 2016, pp. 2–6, 2017.
- W. M. P. Duhita, "Clustering Menggunakan Metode K-Means untuk Menentukan Status Gizi Balita," *J. Inform.*, vol. 15, no. 2, pp. 160–174, 2016.
- A. K. Wardhani, "Implementasi Algoritma K-Means untuk Pengelompokan Penyakit Pasien pada Puskesmas Kajen Pekalongan," *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.
- M. Hariyanto and R. T. Shita, "Penyakit DBD Menggunakan Metode Algoritma K-Means dan Metode," vol. 1, no. 1, pp. 117–122, 2018.
- M. A. Wahyu, "Penerapan Metode K-Means Clustering Untuk Mengelompokkan Potensi Produksi Buah – Buahan di Provinsi Daerah Istimewa Yogyakarta," 2017.
- R. R. Putra and C. Wadisman, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means," vol. 1, no. 1, pp. 72–77, 2018.
- A. F. Khairati, A. A. Adlina, G. F. Hertono, and B. D. Handari, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," *Pros. Semin. Nas. Mat.*, vol. 2, pp. 161–170, 2019.
- R. D. Ramadhani, D. J. Ak, and J. D. I. Panjaitan, "Evaluasi K-Means dan K-Medoids pada Dataset Kecil," no. September, pp. 20–24, 2017.