

# Analisis Algoritma *K-Means Clustering* Untuk Daerah Penyebaran Sampah di Kota Bekasi

Dinda Ruwandara<sup>1</sup>, Mohamad Jajuli<sup>2</sup>, Adhi Rizal<sup>3</sup>

Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang  
e-mail: <sup>1</sup>dinda.16074@student.unsika.ac.id, <sup>2</sup> mohamad.jajuli@unsika.ac.id,  
<sup>3</sup>adhi.rizal@staff.unsika.ac.id.

Diterima: 29 September 2020; Direvisi: 13 Maret 2021; Disetujui: 6 Mei 20201

## **Abstrak**

Sampah sudah menjadi masalah nasional. Masalah sampah terkait erat dengan pertumbuhan populasi, pertumbuhan ekonomi dan perubahan pola konsumsi. Penduduk Kota Bekasi setiap hari menghasilkan 1900 ton sampah ke Tempat Pembuangan Akhir tanpa membutuhkan biaya kecil dan armada yang cukup serta lahan Tempat Pembuangan Akhir yang luas. Fasilitas dan infrastruktur pengolahan sampah adalah kebutuhan utama untuk meringankan masalah sampah. Data mining dapat digunakan dalam menemukan area distribusi sampah. Tujuan dari penelitian ini yaitu mengelompokkan distribusi sampah dengan sampah tinggi, sedang dan rendah di Kota Bekasi. Pengujian dilakukan dengan 3 skenario berbeda dan nilai  $K$  untuk mendapatkan nilai  $K$  optimal. Setelah itu, dilakukan 3 skenario dan nilai benih acak berbeda dilakukan. pengujian dievaluasi menggunakan SSE (Sum of Square Error). Hasil evaluasi menunjukkan bahwa jumlah optimal  $K_3$  dengan selisih sebesar SSE 58.889504 pada tahun 2018. Sedangkan pada 2019 memiliki selisih sebesar SSE 47.15601. Cluster pada tahun 2018 dari tiga nilai random seed ( $S$ ) menghasilkan nilai yang sama dengan nilai SSE 26.03858, 2019 cluster dari tiga nilai  $S$  menghasilkan nilai paling optimal pada  $S = 25$  dengan nilai SSE 33.868462.

**Kata kunci:** Clustering, Data Mining, *K-Means*, Sampah

## **Abstract**

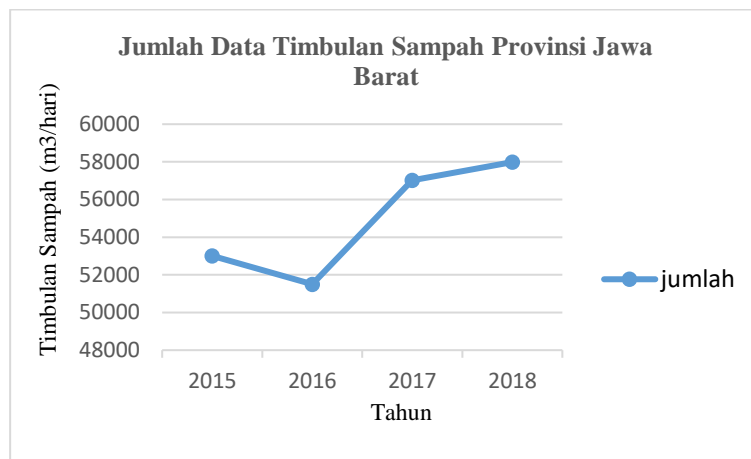
Garbage has become a national problem. Garbage problem is closely related to population growth, economic growth and changes in consumption patterns. Every day residents of Bekasi City produce 1900 tons of garbage to landfills without the need for small costs and sufficient fleets as well as extensive land for landfills. Garbage processing facilities and infrastructure are the main needs to alleviate the garbage problem. Data mining can be used to find waste distribution areas. The purpose of this study is to classify the distribution of high, medium and low waste in Bekasi City. The test was carried out with 3 different scenarios and the  $K$  value to get the optimal  $K$  value. After that, 3 scenarios were carried out and different random seed values were carried out. testing is evaluated using SSE (Sum of Square Error). The evaluation results show that the optimal amount of  $K_3$  is with a difference of SSE 58.889504 in 2018. Meanwhile, in 2019 it has a difference of SSE 47.15601. The cluster in 2018 of three random seed ( $S$ ) values produced the same value as the SSE value 26.03858, 2019 the cluster of three  $S$  values produced the most optimal value at  $S = 25$  with an SSE value of 33.868462.

**Keywords:** Clustering, Data Mining, *K-Means*, Garbage

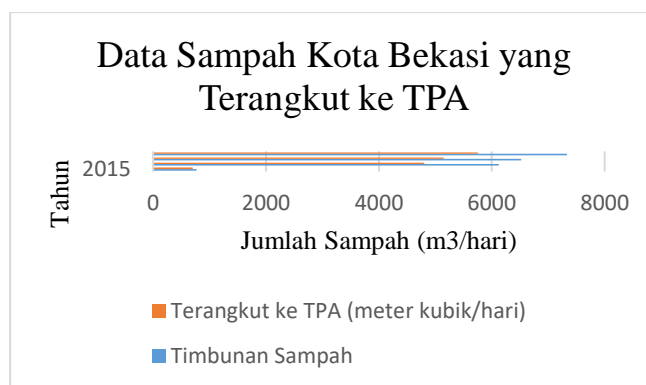
### 1. PENDAHULUAN

Sampah dan limbah telah menjadi permasalahan nasional. Masalah persampahan sangat terkait dengan pertumbuhan penduduk, pertumbuhan ekonomi dan perubahan pola konsumsi masyarakat. Pada tahun 2017 jumlah penduduk Indonesia sudah mencapai 261,89 juta jiwa meningkat dibanding tahun 2000 sebesar 206,26 juta jiwa. Tren pertumbuhan ekonomi juga terus mengalami peningkatan, dengan kontribusi terbesar dari sektor manufaktur. Produk Domestik Bruto yang dihasilkan dari sektor ini sebesar 2.739,4 triliun di 2017, meningkat dari tahun 2000 yang hanya sebesar 385,5 triliun. Pertumbuhan pesat di sektor industri juga merupakan imbas dari meningkatnya pendapatan rumah tangga dan makin beragamnya pola serta jenis konsumsi masyarakat. Kondisi tersebut menimbulkan bertambahnya volume, beragamnya jenis, dan karakteristik sampah dan limbah [1].

Secara keseluruhan sebaran sampah yang paling banyak terdapat di Provinsi Jawa Barat. Hal ini dikarenakan jumlah penduduk di Jawa Barat berada di posisi paling tinggi di Indonesia. Karena faktor penduduk yang terus meningkat dan tidak sebandingnya timbulan sampah yang dihasilkan dengan sampah yang dapat terangkut, sampah di Jawa Barat mengalami peningkatan dari tahun 2015-2018 yang tampak pada gambar 1.



Gambar 1. Jumlah data timbulan sampah provinsi jawa barat. Sumber [2]



Gambar 2. Data sampah kota bekasi yang terangkut ke TPA. Sumber [2]

Sampah-sampah di Provinsi Jawa Barat tersebut tentunya dihasilkan dari Kota atau Kabupaten yang ada di Jawa Barat, salah satunya adalah Kota Bekasi. Pada gambar 2 terlihat jumlah sampah di Kota Bekasi ini terus mengalami peningkatan sejak tahun 2015-2018. Peningkatan tersebut menjadikan Kota Bekasi pada tahun 2018 menduduki posisi pertama dengan timbulan sampah terbanyak yang sebelumnya di ditempati oleh Kota Bandung. Selain itu juga

timbulan sampah ini tidak sebanding dengan sampah yang terangkut setiap harinya, hal ini tentunya akan menyebabkan terjadinya penumpukan sampah di Kota Bekasi.

Sarana dan prasarana pengolahan sampah merupakan suatu kebutuhan utama untuk mengentaskan masalah sampah. Kebijakan pemerintah dalam penanganan sampah di daerah urban harus tepat sasaran. Penambahan jumlah sarana dan prasarana sampah sebaiknya mengikuti tren terhadap jumlah penduduk sehingga permasalahan sampah dapat ditangani dengan baik [3].

Berdasarkan masalah yang telah diuraikan, maka solusi dalam membantu penanganan yang efektif diperlukan suatu metode untuk mengetahui daerah di Kota Bekasi yang menjadi daerah penyebaran timbulan sampah berdasarkan faktor utama ketersediaan sarana dan prasarana yang ada saat ini terhadap timbulan sampah dan jumlah penduduk. Penelitian ini diharapkan menjadi salah satu alat pengambil kebijakan di lingkungan Pemerintah Kota Bekasi. Sehingga, solusi dari permasalahan sampah di Kota Bekasi dapat tertanggulangi dengan baik.

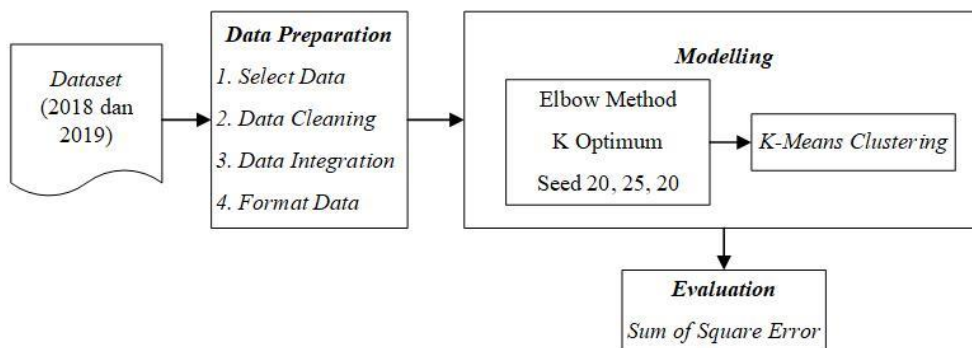
Dalam menemukan daerah penyebaran sampah ini dapat menggunakan Data Mining. Data Mining merupakan suatu proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer untuk menganalisis dan mengekstraksi pengetahuan secara otomatis [4]. Dalam dunia ilmu komputer, Berdasarkan fungsinya, fungsi Data Mining dibagi menjadi 6 bagian yaitu deskripsi, estimasi, prediksi, klasifikasi, *clustering*, dan asosiasi [5]. Salah satu fungsi dari Data Mining adalah *clustering*. *Clustering* adalah mengelompokkan item data ke sejumlah grup kecil yang kemudian masing-masing grup mempunyai kesamaan yang esensial [6]. Adapun algoritma yang dapat digunakan dalam *Clustering* yaitu *K-Means*, *K-Medoids*, *DBSCAN*, dan *Fuzzy C-Means*. Algoritma yang digunakan dalam penelitian ini yaitu algoritma *K-Means*

Berdasarkan penelitian terdahulu bahwa dalam penelitiannya dihasilkan metode *K-Means* memiliki hasil yang lebih baik dari pada metode AHC [7]. Hasil penelitian lain menunjukkan bahwa algoritma *K-Means* dari segi waktu lebih baik dibandingkan algoritma Fuzzy C-Means [8]. serta implementasi algoritma *K-Means* lebih baik daripada implementasi *Fuzzy C-Means* [9].

Berdasarkan permasalahan dan pemaparan penelitian sebelumnya, akan diadakan penelitian mengenai penyebaran sampah di Kota Bekasi serta memanfaatkan data dari Dinas Lingkungan Hidup Kota Bekasi untuk dilakukan klusterisasi menggunakan algoritma *K-Means*. Dengan adanya penelitian ini maka dapat diketahui daerah mana saja yang memiliki timbulan sampah tinggi dan penanganan yang tepat sehingga permasalahan sampah di Kota Bekasi dapat teratasi.

## 2. METODE PENELITIAN

Alur penelitian diilustrasikan pada gambar 3, yang terdiri dari 3 bagian utama yaitu *data preparation*, *modelling*, dan *evaluation*.



Gambar 3. Alur penelitian

2.1. Dataset

Dataset yang didapatkan dari Dinas Lingkungan Hidup Kabupaten Bekasi yaitu dataset timbulan sampah dan dataset sampah yang terangkut. Berikut pada tabel 2 adalah kedua dataset tersebut.

Tabel 1. Deskripsi Dataset

Nama Dataset	Tahun	Atribut	Sumber
Dataset Timbulan Sampah	2018, 2019	Nomor, Nama Kecamatan, Jumlah Penduduk (Jiwa), Produksi Sampah (m <sup>3</sup> /Tahun dan Ton/Tahun)	Dinas Lingkungan Hidup Kabupaten Bekasi
Dataset sampah yang terangkut	2018, 2019	Nomor, Wilayah, Mobil, Jenis Kendaraan, Bulan (Januari-Desember), Ritase, Ritase kg/tonase, Jumlah Tonase, Rata-rata Perbulan	

2.2. Algoritma K-Means

K-Means adalah salah satu metode clustering non hierarki yang berusaha membagi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lainnya [10] Secara umum terdapat langkah-langkah dari K-Means clustering, yaitu [11].

1. Menentukan k sebagai jumlah cluster.
2. Membangkitkan nilai random untuk pusat cluster awal (centroid) sebanyak k.
3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak Euclidean (Euclidean Distance) sehingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Berikut persamaan 1 rumus Euclidean Distance:

$$D(x_i, \pi_i) = \sqrt{\sum_{i=1}^n (x_i - \pi_i)^2} \tag{1}$$

Dimana:

$D(x_i, \pi_i)$  = Jarak antara clustering x dengan pusat cluster  $\mu$  pada data ke-i

$x_i$  = Bobot data ke-i pada cluster yang ingin dicari jaraknya

$\mu_i$  = Bobot data ke-i pada pusat cluster

$n$  = Jumlah data

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil)
5. Memperbaharui nilai centroid. Nilai centroid ini baru diperoleh dari rata-rata cluster yang bersangkutan dengan menggunakan persamaan 2 berikut:

$$C_k = \frac{1}{n_k} \sum_{i=1}^n d_i \tag{2}$$

Dimana:

$n_k$  = jumlah data dalam cluster k

$d_i$  = jumlah dari nilai jarak yang masuk dalam masing-masing cluster

6. Melakukan perulangan dari langkah 2 hingga 5 sampai anggota tiap cluster tidak ada yang berubah. Jika langkah 6 ini sudah terpenuhi, maka nilai pusat cluster pada iterasi terakhir akan digunakan sebagai parameter untuk menemukan klasifikasi data.

### 2.3. Sum of Square Error (SSE)

Sum of Square Error atau SSE dapat digunakan untuk validasi dari masing-masing nilai *cluster*. Karena semakin besar jumlah *cluster* K maka nilai SSE akan semakin kecil. Untuk menghitung SSE menggunakan rumus sebagai berikut [12].

## 3. HASIL DAN PEMBAHASAN

### 3.1. Data Preparation

#### a. Select Data

Data yang diambil mengenai data sampah, yaitu data sampah yang terangkut dan data timbulan sampah di Kota Bekasi. *Dataset* ini disimpan dalam bentuk excel atau \*.csv. Data yang dipilih yaitu data yang berkaitan dengan sampah di tahun 2018 dan 2019. Adapun atribut yang digunakan dalam tabel data timbulan sampah yaitu nama kecamatan, jumlah penduduk, dan jumlah produksi sampah/timbulan sampah. Sedangkan pada tabel sampah yang terangkut data yang digunakan yaitu mobil, ritase (januari-desember), jumlah ritase, dan jumlah sampah terangkut. Pada data tahun 2018 tidak ada atribut sampah yang terangkut, namun sampah terangkut tersebut berada satu kolom dengan kolom ritase. Selain itu juga pada tahun 2019 kolom ritase yang sama dengan tahun 2018 memiliki nama jumlah ritase.

#### b. Data Cleaning

Terdapat beberapa data yang ketika diinputkan kembali pada Microsoft Excel dengan hasil penjumlahan data awal berbeda. Selain adanya perbedaan penjumlahan, terdapat perbedaan atribut dan satuan antara *dataset* tahun 2018 dengan tahun 2019. Dan terdapat beberapa baris yang kosong atau tidak memiliki nilai dalam tabel tersebut. Oleh karena itu untuk *dataset* final menggunakan data yang telah dihitung ulang di Microsoft.Excel, atribut pada Tabel Sampah yang Terangkut Tahun 2018 ditambah yaitu atribut jumlah sampah terangkut, dan jumlah ritase serta menghapus baris yang kosong pada bagian Light Truck. Sedangkan untuk perbedaan satuan dalam Tabel Timbulan Sampah yang digunakan adalah satuan ton/tahun. Dalam *dataset* sampah yang terangkut tahun 2018 terdapat *Missing value* atau data yang hilang. Data tersebut berada di kolom mobil yang wilayahnya menunjukkan wilayah Rawalumbu. Dimana pada bagian *Light Truck* nilai mobil kosong namun memiliki nilai ritase. *Missing value* ini dapat dilihat pada gambar 4.

Wilayah	Mobil	JENIS KENDARAAN			
			Januari	Februari	Maret
MUSTIKAS JAYA	17	ARM & DUMP	349	393	418
	1	LIGHT TRUCK	19	22	23
		TONASE	1.622.070	1.651.570	1.725.830
BANTAR GEBANG	7	ARM & DUMP	127	176	171
		LIGHT TRUCK			
		TONASE	502.390	625.940	652.360
RAWALUMBU	15	ARM & DUMP	199	291	325
		LIGHT TRUCK	20	25	21
		TONASE	1.048.340	1.340.400	1.432.540

Gambar 4. *Dataset Missing value* tabel sampah yang terangkut

#### c. Data Integration

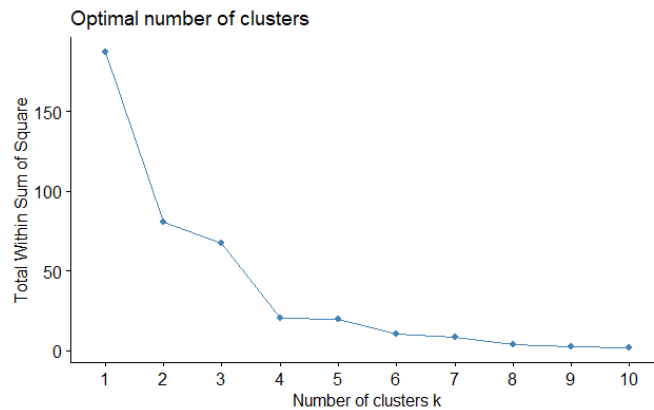
Pada tahap ini dilakukan integrasi data antara *dataset* timbulan sampah dan *dataset* yang terangkut. *Dataset* yang diintegrasikan yaitu *dataset* yang telah dilakukan proses *data cleaning* sehingga menghasilkan satu *dataset* utuh.

d. *Format Data*

Tahapan ini merupakan tahapan terakhir dari pengolahan data. Dalam tahapan ini memproduksi set data akhir yang siap ditambang atau diolah dalam tools pemodelan data mining. Format *dataset* akhir berupa \*.csv untuk pemodelan data mining.

3.2. *Modelling*

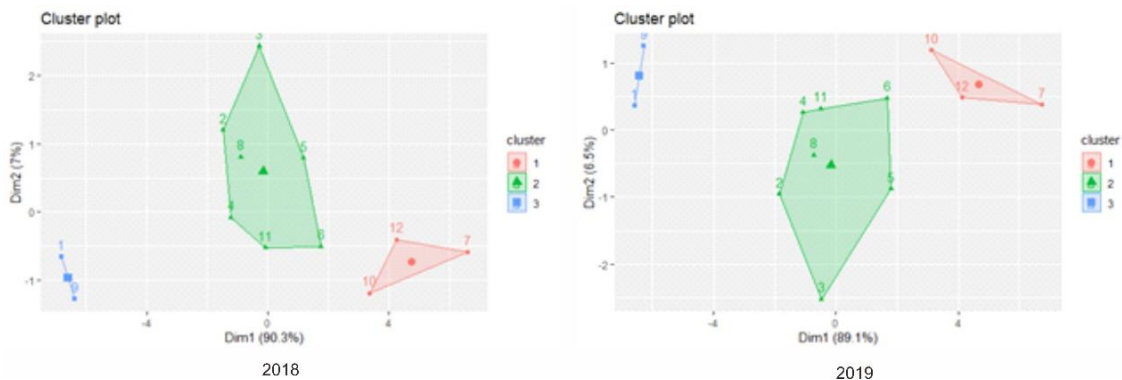
*Dataset* yang telah melalui tahap *data preparation* selanjutnya dilakukan tahap *modelling*. Hal pertama yang dilakukan yaitu menentukan jumlah *cluster* dengan menggunakan metode *Elbow*. Gambar 5 adalah grafik dari jumlah *cluster* optimal.



Gambar 5. Grafik *cluster* optimal dengan metode *Elbow*

Berdasarkan gambar 5, *cluster* optimal yang didapat yaitu 3 *cluster*. Langkah selanjutnya yaitu tahap membangun model dengan menggunakan algoritma *K-Means* dengan kombinasi *seed* yaitu 20, 25, dan 30.

Hasil dari tahap *modelling* ini adalah plot yang menyatakan sebaran sampah di Kota Bekasi. Plot terbaik untuk *dataset* tahun 2018 dan 2019 yaitu dengan *seed* 25. Gambar x merupakan plot dari masing-masing tahun.



Gambar 6. Plot *cluster* tahun 2018 dan 2019

Berdasarkan gambar 6, warna merah mewakili *cluster* 1, warna hijau yaitu *cluster* 2, dan warna biru yaitu *cluster* 3. Pada tahun 2018, *cluster* 1 memiliki 3 anggota, *cluster* 2 memiliki 7 anggota, sedangkan *cluster* 3 memiliki 2 anggota. Pada tahun 2019, *cluster* 1 memiliki 3 anggota, *cluster* 2 memiliki 7 anggota, sedangkan *cluster* 3 memiliki 2 anggota. Tabel 2 adalah hasil dari rata-rata sebaran sampah.

Tabel 2. Anggota *cluster* dan rata-rata sebaran sampah

Tahun	Cluster 1	Cluster 2	Cluster 3
2018	Bantargebang, Jatisampurna, Pondok Melati	Bekasi Barat, Bekasi Utara, Bekasi Selatan, Rawalumbu, Medan Satria, Pondok Gede, Mustika Jaya	Bekasi Timur, Jatiasih
Rata-rata	2502.159	5104.033	6280.191
2019	Bantargebang, Jatisampurna, Pondok Melati	Bekasi Barat, Bekasi Utara, Bekasi Selatan, Rawalumbu, Medan Satria, Pondok Gede, Mustika Jaya	Bekasi Timur, Jatiasih
Rata-rata	2593.284	5311.606	6772.318

Berdasarkan Tabel 2 dapat dilihat bahwa tidak ada perbedaan jumlah anggota dan nama anggota antara tahun 2018 dan 2019, hanya saja terjadi peningkatan rata-rata sebaran sampah dari tahun 2018 ke tahun 2019. Apabila dilihat dari nilai rata-ratanya maka dapat disimpulkan bahwa *cluster* 1 merupakan *cluster* rendah, *cluster* 2 merupakan *cluster* sedang, dan *cluster* 3 merupakan *cluster* tinggi.

### 3.3. Evaluation

Tahap terakhir yaitu evaluasi dengan menggunakan *Sum of Square Error* (SSE). Tabel 3 merupakan nilai SSE.

Tabel 3. Nilai *Sum of Square Error* (SSE)

Tahun	S=20	S=25	S=30
2018	26.03858	26.03858	26.03858
2019	33.98127	33.86462	67.22423

Pada tabel 3 dilihat bahwa pada tahun 2018 memiliki nilai SSE yang sama dalam setiap S. Dengan demikian dapat disimpulkan bahwa *cluster* dengan ketiga *seed* di atas sama baiknya yaitu 26.03858. Pada tahun 2019 *cluster* yang cenderung memiliki nilai SSE yang paling kecil dibandingkan dengan total SSE lain untuk S yang berbeda yaitu K=3 dan S=25 yaitu 33.86462, dengan begitu K=3 dan S=25 merupakan *cluster* terbaik pada tahun 2019.

## 4. KESIMPULAN

Hasil *clustering* perhitungan K=3 dan kombinasi S (*seed*) dikelompokkan pada tahun 2018 dan 2019 jumlah anggota *cluster* 1 yaitu 3 anggota yang dikategorikan daerah penyebaran sampah dalam tingkatan rendah, *cluster* 2 yaitu 7 anggota yang dikategorikan daerah penyebaran sampah dalam tingkatan sedang, lalu *cluster* 3 dengan 2 anggota yang dikategorikan daerah penyebaran sampah dalam tingkatan tinggi.

Hasil *clustering* dievaluasi menggunakan SSE (*Sum of Square Error*) dari K (ukuran *cluster*) dan kombinasi S (*random seed*). Total SSE sudah mencakupi evaluasi untuk *cluster* menggunakan *K-Means* menggunakan nilai yang paling rendah dari keseluruhan percobaan *clustering* pada tahun 2018 dan 2019. Berdasarkan hasil *clustering* daerah penyebaran sampah di Kota Bekasi pada tahun 2018 memiliki hasil SSE yang sama, dengan begitu dalam penelitian ini menggunakan K=3 dan S=25. Sedangkan untuk tahun 2019 hasil *clustering* terbaik adalah K=3 dan S=25, hal ini dikarenakan K=3 dan S=25 di tahun 2019 memiliki nilai SSE paling rendah.

## 5. SARAN

Untuk metode *clustering* daerah penyebaran sampah di Kota Bekasi yang dilakukan pada penelitian ini hanya menggunakan data pada tahun 2018 dan 2019. Maka untuk penelitian selanjutnya diharapkan data yang digunakan lebih banyak. Diharapkan penelitian selanjutnya dapat menambah variabel yang lebih bervariasi lagi. Disarankan agar penelitian selanjutnya menggunakan algoritma *clustering* dan *tools* yang berbeda dalam mengelompokkan daerah penyebaran sampah di Kota Bekasi. Hasil dari pengolahan data disarankan agar dapat diimplementasikan ke dalam program untuk dijadikan suatu aplikasi pengklasteran.

## DAFTAR PUSTAKA

- [1] B.P.S.Indonesia, 2016, Statistik Lingkungan hidup Indonesia, Jakarta.
- [2] BPS Indonesia, 2016, Provinsi Jawa Barat Dalam Angka 2016, vol. 1, Jakarta.
- [3] F.Agus and M.Mahyudin. 2017. Pengklasifikasian Dokumen Berbahasa Arab Menggunakan K-Nearest Neighbor, *Jurnal SIFO Mikroskil*, vol. 18, no. 1, hal 43–56.
- [4] M.G.Sadewo, A.P.Windarto, and D.Hartama. 2017. Penerapan data mining pada populasi daging ayam ras pedaging di indonesia berdasarkan provinsi menggunakan k-means clustering. *InfoTekJar (Jurnal Nasional Informasi dan Teknologi Jaringan)*, vol. 2, no. 1, hal 60–67.
- [5] A.Bastian. 2018. Penerapan algoritma k-means clustering analysis pada penyakit menular manusia (studi kasus kabupaten Majalengka). *Jurnal Sistem Informasi*, vol. 14, no. 1, hal 28–34.
- [6] A.Asrani and R.Adrian. 2016. Penerapan metode K-means untuk clustering mahasiswa berdasarkan nilai akademik dengan Weka Interface studi kasus pada jurusan Teknik Informatika UMM Magelang. *Semesta Teknologi*, vol. 18, no. 1, hal 76–82.
- [7] L.Zahrotun. 2015. Analisis pengelompokan jumlah penumpang bus Trans Jogja menggunakan metode clustering k-means dan agglomerative hierarchical clustering (AHC). *Jurnal Informatika Ahmad Dahlan*, vol. 9, no. 1.
- [8] Yohannes. 2017. Analisis Perbandingan Algoritma Fuzzy C-Means dan K-Means,” in Annual Research Seminar (ARS), 2017, vol. 2, no. 1, pp. 151–155.
- [9] H.Artanto, I.Istiadi, F.Marisa, and D.Purnomo. 2019. Implementasi dan komparasi algoritma fuzzy c-means dan k-means untuk mengelompokkan siswa berdasarkan nilai akademik dan perilaku siswa (data survey), *Conference on Innovation and Application of Science and Technology (CIASTECH)*.
- [10] B.M.Metisen and H.L.Sari. 2015. Analisis clustering menggunakan metode K-Means dalam pengelompokkan penjualan produk pada Swalayan Fadhila. *Jurnal media infotama*, vol. 11, no. 2.
- [11] N.Rohmawati, S.Defiyanti, and M.Jajuli. 2015. Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa,” *Jurnal Ilmu Teknologi Infomasi Terapan*, vol. 1, no. 2,
- [12] A.T.Rahman. 2017. Coal Trade Data Clustering Using K-Means (Case Study Pt. Global Bangkit Utama). *ITSMART Jurnal Teknologi dan Informasi*, Vol. 6 No. 1 hal 24-31