Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa

Elly Muningsih[1] , Ina Maryani[2] , Vembria Rose Handayani[3]
[1,3] Universitas Bina Sarana Informatika
[2] Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri

elly.emh@bsi.ac.id[1], ina.maryani@nusamandiri.ac.id[2] , vembria.vrh@bsi.ac.id[3]

Abstrak : Metode K-Means merupakan salah satu metode Data Mining yang banyak digunakan dalam penelitian pengelompokan. Namun metode K-Means memiliki beberapa kekurangan, salah satunya yaitu dalam penentuan jumlah cluster. Penelitian kali ini akan mengaplikasikan Indeks Davies Bouldin (DBI) sebagai salah satu cara optimasi jumlah cluster untuk mengelompokkan propinsi berdasar potensi desa dengan banyaknya jenis industri yang dimiliki wilayahnya. Data yang digunakan adalah data banyaknya desa atau kelurahan menurut keberadaan dan jenis industri kecil dan mikro (desa). Pengolahan data menggunakan aplikasi RapidMiner. Pengujian dilakukan dengan mencari nilai terkecil dari DBI dimana setelah data di olah diketahui nilai terkecil adalah 0,175 di jumlah cluster 3.

Kata kunci : metode K-Means, pengelompokan, Index Davies Bouldin,

Abstract: The K-Means method is one of the most widely used data mining methods in clustering research. However, the K-Means method has several shortcomings, one of which is in determining the number of clusters. This research will apply the Davies Bouldin Index (DBI) as a way of optimizing the number of clusters to classify provinces based on village potential and the number of types of industry the region has. The data used is data on the number of villages or sub-districts according to the existence and type of small and micro industries (villages). Data processing uses the Rapid Miner application. Testing is done by finding the smallest value from the DBI where after the data is processed it is known that the smallest value is 0.175 in the number of clusters 3.

Keywords: K-Means method, grouping, Davies Bouldin Index,

1. Pendahuluan

Potensi Desa atau Podes sejak tahun 1980 sudah dilakukan pendataan oleh Badan Pusat Statistik (BPS). Sejak saat itu, Podes dilaksanakan secara rutin sebanyak 3 kali dalam kurun waktu sepuluh tahun untuk mendukung kegiatan Sensus Penduduk, Sensus Pertanian, ataupun Sensus Ekonomi. Dengan demikian, fakta penting terkait ketersediaan infrastruktur dan potensi yang dimiliki oleh setiap wilayah dapat dipantau perkembangannya secara berkala dan terus menerus. Salah satu potensi desa yang dimaksud adalah usaha industri. Usaha industri adalah suatu unit (kesatuan) usaha yang melakukan kegiatan ekonomi, bertujuan menghasilkan barang atau jasa, terletak pada suatu bangunan atau lokasi tertentu, dan mempunyai catatan administrasi tersendiri mengenai produksi dan struktur biaya serta ada seorang atau lebih yang bertanggung jawab atas usaha tersebut (www.bps.go.id). Penggolongan perusahaan industri pengolahan ini semata-mata hanya didasarkan kepada banyaknya tenaga kerja yang bekerja, tanpa memperhatikan apakah perusahaan itu menggunakan mesin tenaga

atau tidak, serta tanpa memperhatikan besarnya modal perusahaan itu.

Klasifikasi industri yang digunakan dalam survei industri pengolahan adalah klasifikasi yang berdasar kepada *International Standard Industrial Classification of all Economic Activities (ISIC) revisi 4*, yang telah disesuaikan dengan kondisi di Indonesia dengan nama Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) tahun 2009. Klasifikasi industri yang dimaksud antara lain adalah makanan, minuman, pengolahan tembakau, tekstil, pakaian jadi, kulit, kayu, kertas, pencetakan dan reproduksi media rekaman, Produk dari batu bara dan pengilangan minyak bumi, bahan kimia dan barang dari bahan kimia, farmasi, produk obat kimia dan obat tradisional, karet, barang dari karet dan plastik, barang galian bukan logam, logam dasar, barang logam, bukan mesin dan peralatannya, komputer, barang elektronik dan dan optic, peralatan listrik, mesin dan perlengkapan, kendaraan bermotor, trailer dan semi trailer, alat angkutan lainnya, furniture, pengolahan lainnya, jasa reparasi dan pemasangan mesin dan peralatan.

Penelitian ini akan mengelompokkan propinsi di Indonesia berdasarkan jumlah desa atau kelurahan yang memiliki industri sebagai bagian dari Podes dari wilayah yang bersangkutan. Metode yang digunakan adalah metode clustering K-Means yang merupakan metode pengelompokan yang sering digunakan dalam penelitian. Metode K-Means merupakan salah satu metode clustering yang ada di Data Mining. Jumlah kelompok ditentukan dari evaluasi dilakukan dengan mencari nilai terkecil dari nilai Indeks Davies Bouldin sebagai cara untuk optimasi jumlah cluster.

Analisis cluster merupakan suatu proses pemisahan objek menjadi beberapa kelopok sehingga objek yang masuk dalam kelompok yang sama memiliki karkteristik atau kemiripan yang sama dan akan berbeda dengan objek lain pada kelompok lainnya (Dewi Kusumah, Warsito, & Abdul Mukid, 2017). Clustering atau pengelompokan merupakan salah satu dari metode Data Mining yang membagi data ke dalam beberapa kelompok dimana objek dengan kemiripan atau karakteristik samaakan menjadi satu kelompok (Irhamni, Damayanti, Khusnul K, & A, 2014). Clustering data digunakan untuk membagi data menjadi beberapa kelompok berdasar kemiripan pola yang sama (Irhamni et al., 2014). Metode clustering adalah suatu metodeyang digunakan dalam pengelompokan suatu himpunan data menjadi beberapa beberapa kelompok atau klaster sehingga data dalam satu klaster memiliki karakteristik dan kemiripan yang sama, sedangkan data dalam klaster yang berbeda memiliki karakteristik yang berbeda pula(Nanda, Mahanty, & Tiwari, 2010).

Salah satu metode clustering yang digunakan dalam penelitian pengelompokan data adalah metode K-Means (Muningsih, 2018. Metode K-Means adalah metode sederhana untuk membagi suatu kumpulan atau himpunan data dalam suatu angka spesifik dari sebuah cluster, yaitu nilai (Larose & Larose, 2014). K-Means merupakan suatu metode data clustering non hirarki yang mempartisi data ke dalam bentuk satu atau lebih cluster atau kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam kelompok yang lain (Yudi Agusta, 2007). Algoritma metode K-Means untuk clustering dapat dilakukan dengan cara (Yudi Agusta, 2007), (Muningsih & Kiswati, 2015) :

1. Tentukan banyaknya atau jumlah cluster yang akan dibentuk
2. Inisialisasi nilai k sebagai pusat dari cluster (beri nilai random)
3. Alokasikan data yang diolah sesuai dengan jumlah cluster yang sudah ditentukan. Kedekatan dari dua obyek ditentukan oleh jarak antar kedua obyek tersebut. Jarak yang paling

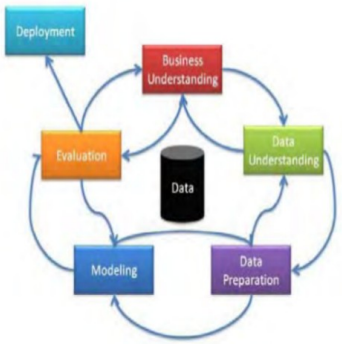
dekat antara satu data dengan satu cluster lain akan menentukan suatu data masuk ke dalam cluster yang mana.

4. Hitung nilai centroid (pusat cluster) pada tiap dari cluster. Pusat cluster merupakan rata-ratasemua data atau obyek dalam sebuah cluster.
5. Alokasikan lagi setiap obyek menggunakan pusat cluster yang baru. Jika nilai pusat cluster sudah tetap, tidak berubah lagi maka proses pengelompokan atau peng-clusteran selesai.
6. Kembali lagi ke langkah 3 sampai pusat dari cluster tidak berubah lagi

Evaluasi clustering dilakukan dengan tujuan untuk mengetahui seberapa baik kualitas dari hasil clustering. Pada penelitian ini, evaluasi hasil clustering yang digunakan adalah Davies Bouldin Index untuk mengetahui jumlah cluster yang paling optimal. Davies Bouldin Index (DBI) diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979 Davies-Bouldin Index merupakan salah satu metode yang digunakan untuk mengukur validitas ataujumlah cluster paling optimal pada suatu metode pengelompokan dimana kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari cluster yang diikuti (Bates & Kalita, 2016). Evaluasi dengan menggunakan Davies Bouldin Index ini memiliki skema evaluasi dari internal cluster, dimana baik atau tidaknya hasil cluster dilihat dari kuantitas dan kedekatan antar data hasil cluster (Bates & Kalita, 2016).

2. Metode Penelitian

Untuk membangun model pada penelitian ini digunakan metode CRISP-DM (Cross-Industry Standard Process for Data Mining). Metode ini memiliki enam fase atau tahapan seperti yang ditampilkan pada gambar 1 :



Sumber : (Sastry & Babu, 2013)

Gambar 1. Tahapan pada CRISP-DM

Pada penelitian ini, tahapan yang dilakukan adalah mengumpulkan dataset, memilih atribut yang relevan, membangun model clustering menggunakan metode clustering K-Means untuk pengelompokan data dan evaluasi model dengan Index Davies Bouldin.

2.1. Dataset

Dataset yang digunakan adalah data Banyaknya Desa/Kelurahan Menurut Keberadaan dan Jenis Industri Kecil dan Mikro (Desa) tahun 2018 yang diambil dari website BPS (www.bps.go.id) dengan jumlah data sebanyak jumlah propinsi di Indonesia dan atribut berjumlah 8. Record berisi nama-nama propinsi di Indonesia dan atribut jenis industri sebagai potensi desa. Informasi lengkap data banyaknya desa dan potensi industrinya yang digunakan adalah :

1. Propinsi : Aceh, Sumatera Utara, Sumatera Barat, Riau dan seterusnya.
2. Data jenis industri : Industri Makanan dan Minuman, Industri dari Kain/Tenun, Industri Gerabah, Keramik/Batu, Industri Anyaman, Industri Logam Mulia dan Bahan dari Logam dan lain sebagainya.

2.2. Preprocessing Data

Dari dataset yang ada kemudian dilakukan reprocessing data, salah satunya yaitu mengubah nama atribut jenis industry menjadi kode tertentu agar lebih mudah untuk diolah yaitu I1 sampai I8. Tipe data untuk atribut yang digunakan adalah

integer. Data lengkap yang diolah ditampilkan pada gambar 2 berikut ini :

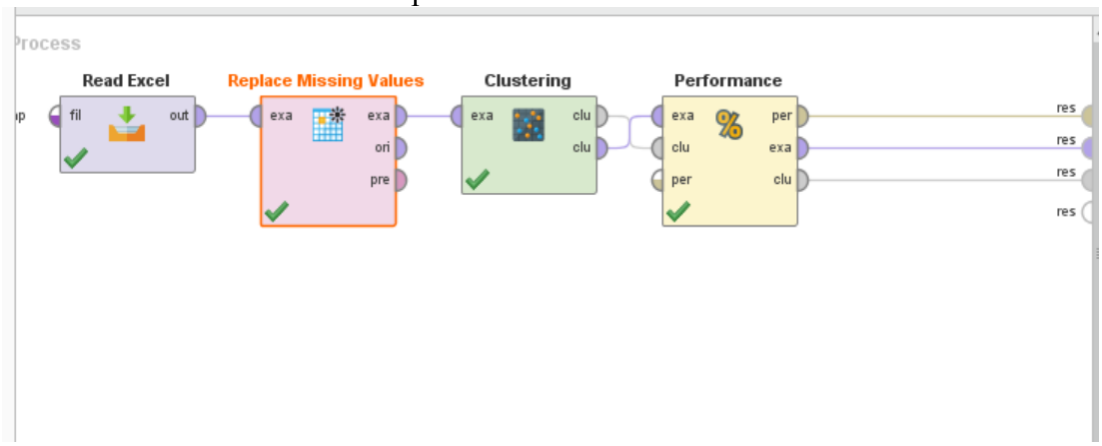
Propinsi	I1	I2	I3	I4	I5	I6	I7	I8
ACEH	1424	504	368	517	320	1146	25	310
SUMATERA UTARA	1623	1004	362	679	370	1225	87	524
SUMATERA BARAT	897	582	433	347	279	944	120	299
RIAU	548	303	251	293	151	781	18	127
JAMBI	443	188	154	194	139	720	7	111
SUMATERA SELATAN	843	315	363	411	207	1179	19	247
BENGKULU	538	179	174	216	108	527	7	162
LAMPUNG	1076	572	714	550	253	1432	25	278
KEP. BANGKA BELITUNG	276	76	154	160	77	258	1	34
KEP. RIAU	305	109	92	114	79	219	26	115

Sumber : Penulis (2021)

Gambar 2. Preprocessing Data

2.3. Modelling dan Evaluasi

Pada tahapan ini, dilakukan modelling menggunakan tools RapidMiner dengan metode K-Means. Model yang digunakan menambahkan operator Replace Missing Values untuk menghapus data-data yang memiliki nilai miss atau null dan operator Performance untuk menghitung nilai Index Davies Bouldin (IDB). Nilai terkecil dari IDB menunjukkan jumlah cluster paling optimal. Gambar 3 menampilkan Modelling dan Evaluasi menggunakan tools RapidMiner.



Sumber : Penulis (2021)

Gambar 3. Modelling dan Evaluasi

Langkah dan tahapan dalam proses modelling dan evaluasi adalah :

a. Membuat model clustering dengan metode K-Means dimana jumlah cluster yang dimodel adalah 2 – 10. Metode clustering menggunakan metode K-Means
b. Dari tiap cluster yang dibuat di evaluasi dengan operator Cluster Distance Performance untuk mengetahui nilai DBI tiap cluster.
c. Nilai DBI yang terkecil menunjukkan hasil yang paling baik dan menunjukkan jumlah cluster yang optimal.

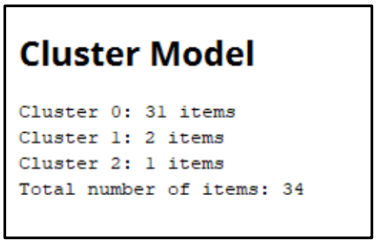
3. Hasil dan Pembahasan

Proses pengolahan data yang dilakukan dengan modeling menggunakan Metode K-Means, dimana dicari nilai DBI terkecil untuk mengetahui optimasi jumlah clusternya. Jumlah cluster dan nilai DBI-nya ditampilkan pada Tabel 1 berikut ini :

Tabel 1. Nilai DBI Tiap Cluster

Cluster	Nilai DBI
2	0,309
3	**0,175**
4	0,423
5	0,575
6	0,720
7	0,419
8	0,501
9	0,604
10	0,626

Dari hasil pengolahan diketahui, untuk optimasi jumlah cluster adalah 3 dengan nilai DBI 0,175. Maka untuk data ini, pengelompokan yang dilakukan dengan optimasi jumlah cluster adalah 3. Jumlah anggota masing-masing cluster ditampilkan pada gambar 4 dibawah ini :



Sumber : Penulis (2021)

Gambar 4. Jumlah Anggota Tiap Cluster

Dari gambar diatas dapat dijelaskan bahwa untuk cluster 0 memiliki anggota 31 propinsi, cluster 1 memiliki anggota 2 propinsi dan cluster 2 memiliki anggota 1 propinsi.

Dan untuk nilai centroid masing-masing cluster ditampilkan pada Gambar 5 seperti dibawah ini :

Attribute	cluster_0	cluster_1
I1	680.710	5817.500
I2	322.258	3288.500
I3	269.742	2612
I4	323.484	2629
I5	147.290	1273.500
I6	691.581	6275.500
I7	32.935	718
I8	200.290	1640

Sumber : Penulis (2021)

Gambar 5. Nilai Centroid

Dari nilai dan tabel diatas diketahui bahwa tiap cluster untuk kategori propinsi berdasarkan potensi daerah yaitu industri yang ada dibedakan menjadi :

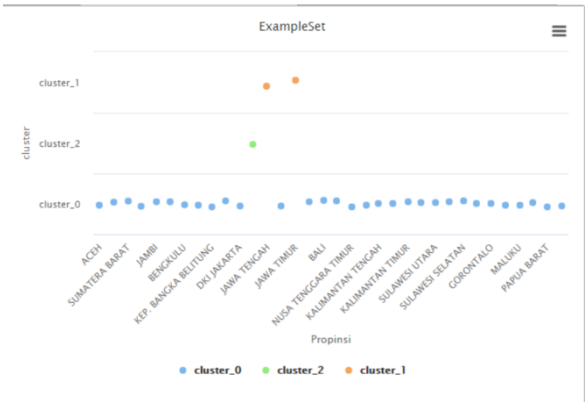
1. Cluster 0 : propinsi dengan potensi desa (industri) sedikit
2. Cluster 1 : propinsi dengan potensi desa (industri) banyak
3. Cluster 2 : propinsi dengan potensi desa (industri)sedang

Untuk data propinsi tiap cluster ditampilan pada Tabel 2 berikut ini :

Tabel 2. Data Propinsi per cluster

Cluster	Jumlah Anggota	Propinsi
0	31	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep, Bangka Belitung, Kep. Riau, DKI Jakarta dan lain sebagainya.
1	2	Jawa Tengah, Jawa Timur
2	1	Jawa Barat

Dari data diatas diketahui bahwa, Jawa Tengah dan Jawa Timur menjadi Ppropinsi yang memiliki industri sebagai potensi desa yang terbesar atau terbanyak. Jawa Barat menjadi satu-satunya Propinsi dengan kategori memiliki jumlah industri sedang. Dan ada 31 Propinsi lainnya masuk kategori sedikit untuk jumlah industri sebagai potensi desanya. Dari hasil clustering kemudian ditampilkan grafik penyebaran anggota cluster pada gambar 6 dibawah ini :



Sumber : Penulis (2021)

Gambar 6. Penyebaran Anggota Cluster

4. Kesimpulan

Dari pengolahan data yang sudah dilakukan, didapatkan hasil dan kesimpulan bahwa modelling K-Means dengan evaluasi nilai DBI menghasilkan optimasi jumlah cluster. Hasil clustering menghasilkan pengelompokan propinsi berdasarkan jumlah industri sebagai potensi desanya dengan kategori sedikit, sedang dan banyak. Karena keterbatasan waktu dan tenaga, Peneliti menyadari bahwa hasil dari penelitian ini masih jauh dari kata sempurna karena hal tersebut maka untuk penelitian berikutnya bisa dilakukan komparasi dengan metode clustering yang lain.

Referensi

Bates, A., & Kalita, J. (2016). Counting clusters in twitter posts. *ACM International Conference Proceeding Series*, *04-05-March-2016*. https://doi.org/10.1145/2905055.2905295

Dewi Kusumah, R., Warsito, B., & Abdul Mukid, M. (2017). Perbandingan metode k – means dan self organizing map (Studi kasus: pengelompokan kabupaten/kota di jawa tengah berdasarkan indikator indeks pembangunan manusia 2015). *Jurnal Gaussian, Vol 6 No 3 Tahun 2017*, *6*, 429–437.

Irhamni, F., Damayanti, F., Khusnul K, B., & A, M. (2014). Optimalisasi pengelompokan kecamatan berdasarkan indikator pendidikan menggunakan metode clustering dan davies bouldin index. *Seminar Nasional Dan Teknologi UMJ*, (11), 1–6.

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908747). https://doi.org/10.1002/9781118874059

Muningsih, E. (2018). Komparasi Metode Clustering K-Means dan K-Medoids dengan Model Fuzzy RFM untuk Pengelompokan Pelanggan. *JurnalEvolusi,6*(2)

Muningsih, E., & Kiswati, S. (2015). Penerapan Metode K-Means Untuk Clustering Produk Online Shop. *Jurnal Bianglala Informatika*, *3*(1).

Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Expert Systems with Applications Clustering Indian stock market data for portfolio management. *Expert Systems With Applications*, *37*(12), 8793–8798. https://doi.org/10.1016/j.eswa.2010.06.026

Yudi Agusta. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem Dan Informatika*, *3*(Februari), 47–60.