BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Clustering merupakan proses mengelompokan data dalam *dataset* menjadi kelompok-kelompok atau *cluster-cluster* yang menggambarkan suatu keadaan tertentu dengan prinsip objek dalam satu *cluster* memiliki kemiripan tinggi tetapi sangat berbeda dari objek dalam *cluster* lainnya (Shakeel dkk., 2018; Han dkk., 2012). Metode Partisi merupakan salah satu metode *clustering* berbasis *mean* atau *medoid* untuk mempresentasikan pusat *cluster (centroid)* dengan satu tingkat pengelompokan. Ditinjau dari karakteristik, kemudahan implementasi dan performa komputasi untuk pengelompokan data berukuran kecil sampai menengah metode partisi berbasis *mean* adalah yang paling efektif dan efisien (Wilks, 2011).

Metode *Clustering* berbasis partisi dengan menggunakan algoritma *K-Means* telah banyak diimplementasikan di berbagai bidang, sebagai contoh di bidang pendidikan *K-Means clustering* digunakan salah satunya untuk pemetaan sebaran guru Sekolah Menengah Atas (SMA) di Indonesia dengan cara mengelompokkan daerah-daerah di Indonesia yang kekurangan dan kelebihan Guru menggunakan rasio siswa terhadap Guru sebagai tolok ukur kebutuhan Guru sehingga diketahui daerah-daerah mana saja yang diutamakan untuk dipenuhi permintaan kebutuhan Gurunya (Widiyaningtyas dkk., 2017). Di bidang teknologi informasi khususnya teknologi jaringan *K-Means clustering* diimplementasikan pada proses pembuatan sub-kelompok virtual dari *node* sensor, yang membantu *node* sensor untuk menurunkan perhitungan routing dan untuk menurunkan ukuran data routing (Purnima dan Arvind, 2014). Pada penelitian tersebut menggunakan metode *Elbow* untuk menentukan jumlah *cluster* terbaik yang digunakan sehingga didapatkan kualitas hasil *cluster* terbaik untuk meningkatkan efektifitas dan efisiensi energi yang digunakan pada jaringan Wireless Sensor Network (WSN). Implementasi *K-Means clustering* dengan optimasi penentuan jumlah *cluster* terbaik menggunakan metode *Elbow* di bidang bisnis juga digunakan untuk mengelompokkan data profil dari pelangan produk batik dengan berbagai jumlah

data percobaan sedangkan jumlah *cluster* yang digunakan antara dua sampai dengan delapan *cluster*. Hasil yang didapatkan pada penelitian tersebut menunjukkan bahwa optimasi algoritma *K-Means* pada penentuan jumlah *cluster* terbaik menggunakan metode *Elbow* menghasilkan jumlah *cluster* yang sama pada jumlah data yang berbeda-beda (Syakur dkk., 2018).

Metode Elbow merupakan metode visualisasi selisih *Within Sum of Square Error* (WSSE) yang membentuk sudut siku pada grafik. Evaluasi beberapa metode optimasi jumlah *cluster* pada algoritma *K-Means* telah dilakukan dengan hasil Metode *Elbow* memiliki keunggulan waktu komputasi lebih efisien dibandingkan dengan metode *Gap Statistic*, metode *Silhouette Coefficient,* metode *Canopy* jika diterapkan pada *dataset* yang kecil dengan hasil jumlah *cluster* yang dihasilkan sama, namun jika yang digunakan adalah *dataset* besar maka metode *Canopy* adalah yang paling efisien (Yuan dan Yang, 2019). Berdasarkan penelitian-penelitian tersebut dapat dikatakan penentuan jumlah *cluster* terbaik menggunakan metode *Elbow* dapat meningkatkan efektifitas dan efisiensi algoritma *K-Means* untuk menghasilkan *cluster* yang baik.

Optimasi algoritma *K-Means clustering* tidak hanya pada penentuan jumlah *cluster* terbaik saja tetapi penentuan *centroid* awal juga sangat menentukan anggota *cluster* yang dihasilkan dan dapat mempengaruhi kinerja algoritma *K-Means* dalam melakukan *clustering*. Penelitian terkait optimasi pada penentuan *centroid* awal salah satunya adalah berbasis nilai *median*, dalam penelitian tersebut data dibagi menjadi sejumlah *cluster* yang telah ditentukan kemudian *centroid* awal ditentukan dengan menggunakan rumus *median* pada masing-masing kelompok data, langkah selanjutnya sesuai dengan algoritma *K-Means* pada umumnya. Hasil percobaan dengan perbandingan metode *K-Means* standart menggunakan *centroid* awal secara *random*, metode *K-Medoid*, metode *Minmax K-Means* menunjukkan bahwa hasil *Davies Bouldin Index (*DBI*)* untuk *K-Means* dengan *centroid* awal berbasis *median* paling kecil, semakin kecil DBI semakin baik kualitas *cluster* yang dihasilkan (Sopian, 2017).

Penggunaan nilai hasil rumus kuartil tengah (median) yang mewakili nilai tengah data untuk penentuan *centroid* awal dibandingkan dengan penentuan

centroid awal secara *random* pernah digunakan pada analisis perbandingan algoritma *K-Means* dan *K-Median* untuk menentukan alokasi bantuan dana pendidikan Provinsi Jawa Tengah, hasil *clustering* pada penelitian tersebut menerangkan bahwa penggunaan nilai kuartil tengah untuk *centroid* awal pada kedua algoritma menghasilkan anggota *cluster* dan jumlah iterasi yang sama, namun penentuan *centroid* awal secara *random* menghasilkan anggota *cluster* dan jumlah iterasi yang berbeda-beda atau bervariasi dengan kata lain pengunaan nilai kuartil tengah (median) untuk menentukan *centroid* awal memiliki hasil yang lebih baik jika dibandingkan penentuan *centroid* awal secara *random*.

2.2. Dasar Teori

2.2.1. Aturan Kepegawaian Sumber Daya Aparatur Guru

a. Rasio Kebutuhan Guru

Guru merupakan salah satu bagian dasar dari sistem pendidikan karena memiliki peran penting dan menentukan dalam kualitas pendidikan salah satunya adalah interaksi antara guru dengan siswa sangat menentukan keberhasilan akademik. Hal tersebut tidak lepas dari rasio murid terhadap guru yang menggambarkan jumlah rata-rata siswa yang diajar oleh guru di sekolah. Jika jumlah rata-rata siswa per guru menurun maka prestasi akademik akan naik (Koc dan Celik, 2015) namun demikian rasio tersebut tidak boleh terlalu tinggi atau rendah sehingga menjadikannya tidak efektif dari segi kuantitas dan pembiayaan.

Rasio murid-guru (*Student-Teacher Ratio*, STR) sekolah-sekolah di Indonesia jauh lebih rendah dibanding dengan negara-negara lain dengan tren menurun, hal ini dikawatirkan terkait dengan efisiensi sistem pendidikan di Indonesia (The World Bank, 2011). Di Indonesia rasio siswa-guru ditentukan dengan Peraturan Pemerintah Republik Indonesia Nomor 74 TH 2008 Tentang Guru antara lain:

1. TK, RA, atau yang sederajat 15:1;
2. SD, SMP atau yang sederajat 20:1;
3. MI, MTs atau yang sederajat 15:1;
4. SMA atau yang sederajat 20:1;

5. MA atau yang sederajat 15:1;
6. SMK atau yang sederajat 15:1;
7. MAK atau yang sederajat 12:1.

Jumlah peserta didik dalam satu rombongan belajar diatur dalam Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 17 Tahun 2017 Tentang Penerimaan Peserta Didik Baru Pada Taman Kanak-Kanak, Sekolah Dasar, Sekolah Menengah Pertama, Sekolah Menengah Atas, Sekolah Menengah Kejuruan, Atau Bentuk Lain Yang Sederajat sebagai berikut:

1. SD dalam satu kelas berjumlah paling sedikit 20 (dua puluh) peserta didik dan paling banyak 28 (dua puluh delapan) peserta didik;
2. SMP dalam satu kelas berjumlah paling sedikit 20 (dua puluh) peserta didik dan paling banyak 32 (tiga puluh dua) peserta didik;
3. SMA dalam satu kelas berjumlah paling sedikit 20 (dua puluh) peserta didik dan paling banyak 36 (tiga puluh enam) peserta didik;
4. SMK dalam satu kelas berjumlah paling sedikit 15 (lima belas) peserta didik dan paling banyak 36 (tiga puluh enam) peserta didik.

b. Kualifikasi Akademik dan Sertifikasi Kompetensi Guru

Peraturan Pemerintah Republik Indonesia Nomor 74 TH 2008 Tentang Guru mewajibkan Guru memiliki kualifikasi akademik, kompetensi, sertifikat pendidik, sehat jasmani dan rohani, serta memiliki kemampuan untuk mewujudkan tujuan pendidikan nasional. Kualifikasi akademik Guru ditunjukkan dengan ijazah yang merefleksikan kemampuan yang dipersyaratkan bagi Guru untuk melaksanakan tugas sebagai pendidik pada jenjang, jenis, dan satuan pendidikan atau mata pelajaran yang diampunya sesuai dengan standar nasional pendidikan, Ijazah tersebut diperoleh melalui pendidikan tinggi program S-1 atau program D-IV pada perguruan tinggi yang menyelenggarakan program pendidikan tenaga kependidikan dan/atau program pendidikan non kependidikan. Kompetensi Guru merupakan seperangkat pengetahuan, keterampilan, dan perilaku yang harus dimiliki, dihayati, dikuasai, dan diaktualisasikan oleh Guru dalam melaksanakan tugas keprofesionalan sedangkan kompetensi Guru meliputi kompetensi pedagogik, kompetensi kepribadian, kompetensi sosial, dan kompetensi profesional yang

diperoleh melalui pendidikan profesi, kompetensi guru ini dibuktikan dengan sertifikat dari uji kompetensi yang diselenggarakan oleh instansi yang berwenang.

Sertifikat Pendidik bagi Guru diperoleh melalui program pendidikan profesi yang diselenggarakan oleh perguruan tinggi yang memiliki program pengadaan tenaga kependidikan yang terakreditasi, baik yang diselenggarakan oleh Pemerintah maupun Masyarakat, dan ditetapkan oleh Pemerintah, Program pendidikan profesi sebagaimana dimaksud hanya diikuti oleh peserta didik yang telah memiliki Kualifikasi Akademik S-1 atau D-IV sesuai dengan ketentuan peraturan perundang-undangan. Berdasarkan data guru dari Dinas Pendidikan dan Kebudayaan Pemerintah Provinsi Jawa Tengah pendikan minimal guru yang mengajar yang di SMA Negeri telah memenuhi kualifikasi minimal berpendidikan S-1 dan lebih dari 90% telah bersertifikasi.

c. Beban Kerja Guru

Beban kerja Guru sesuai dengan Peraturan Pemerintah Republik Indonesia Nomor 74 TH 2008 Tentang Guru mencakup kegiatan pokok merencanakan pembelajaaran, melaksanakan pembelajaran, menilai hasil pembelajaran, membimbing dan melatih peserta didik dan melaksanakan tugas tambahan yang melekat pada pelaksanaan kegiatan pokok sesuai dengan beban kerja Guru. Beban kerja Guru paling sedikit memenuhi 24 (dua puluh empat) jam tatap muka dan paling banyak 40 (empat puluh) jam tatap muka per minggu pada satu atau lebih satuan pendidikan yang memiliki ijin pendirian dari Pemerintah atau Pemerintah Daerah.

Kebutuhan Guru mata pelajaran dihitung berdasarkan Peraturan Bersama Menteri Negara Pendayagunaan Aparatur Negara dan Reformasi Birokrasi, Menteri Pendidikan Nasional, Menteri Dalam Negeri, Menteri Keuangan, dan Menteri Agama Nomor 05/X/PB/2011, SPB/03/M.PAN-RB/10/2011, 48 Tahun 2011, 158/PMK.01/2011, 11 Tentang Penataan dan Pemerataan Guru Pegawai Negeri dengan jumlah jam tatap muka perminggu per jenis guru dikali jumlah rombongan belajar pada tiap tingkat dibagi wajib belajar per minggu yaitu dua puluh empat jam.

d. Pengangkatan, Penempatan, Dan Pemindahan Guru

Menurut Peraturan Bersama Menteri Pendidikan Nasional, Menpan RB, Mendagri, Menteri Keuangan dan Menteri Agama tahun 2011, Guru diangkat oleh Pemerintah dan atau Pemerintah Daerah sesuai dengan aturan peraturan perundang-undangan dengan memperhatikan koordinasi perencanaan kebutuhan Guru secara nasional dan mempertimbangankan pemerataan Guru antar satuan pendidikan yang diselenggarakan Pemerintah Daerah dan atau masyarakat, antar kabupaten atau antar kota, dan anatar provinsi termasuk kebutuhan Guru di Daerah Khusus.

2.2.2. *Data Mining*

Penambangan data atau *data mining* dapat dilihat sebagai hasil dari evolusi alami teknologi informasi (Han dkk., 2012), penambangan data adalah aktivitas mengeksplorasi data untuk mendapatkan pengetahuan yang bermanfaat (Lee dan Ho Kim, 2009) dengan menganalisis pola data pada *database*, istilah data *mining* popular dengan nama *Knowledge Discovery From Data* (KDD), penambangan data merupakan langkah penting dalam proses KDD (Han dkk., 2012). KDD memiliki tahapan sebagai berikut:

a. Pembersihan data

   Untuk menghilangkan noise dan data yang tidak konsisten.

b. Integrasi data

   Banyak sumber data dapat digabungkan.

c. Pemilihan data

   Penggunaan data yang relevan dengan tugas analisis diambil dari basis data.

d. Transformasi data

   Data ditransformasikan dan dikonsolidasikan ke dalam bentuk sesuai untuk penambangan dengan melakukan operasi ringkasan atau agregasi. Pada penelitian ini data ditransformasi dalam skala 0 sampai dengan 1 menggunakan rumus *Min Max Normalization* sebagai berikut (Jain dkk., 2018):

$$X^* = \frac{x - minValue}{max\,Value - minValue} \tag{2.1}$$

dengan X^* adalah nilai normalisasi dan x adalah nilai objek data yang akan dinormalisasi

e. Data mining

Proses penting di mana metode cerdas diterapkan untuk mengekstrak pola data.

f. Evaluasi pola

Mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan pada langkah-langkah tertentu.

g. Presentasi pengetahuan

Teknik visualisasi dan representasi pengetahuan, digunakan untuk menyajikan pengetahuan yang ditambang kepada pengguna.

Fungsi dasar dari proses penambangan data meliputi fitur pemilihan, peringkasan, asosiasi, pengelompokan, prediksi, dan klasifikasi.

2.2.3. *Clustering*

Clustering adalah proses menemukan kelompok (*cluster*) yang berisi pengetahuan dan dapat mengambarkan suatu keaadan tertentu dalam suatu dataset dengan cara mengelompokkan satu set objek data ke dalam beberapa *cluster* berdasarkan kemiripan objek data sehingga objek dalam sebuah *cluster* memiliki kesamaan tinggi, tetapi sangat berbeda dengan objek di *cluster* lain (Kotu dan Deshpande, 2019). Jumlah *cluster* merupakan latar belakang dari pengetahuan yang dicari, tanpa adanya proses *clustering* akan sulit untuk mengiterpretasikan suatu dataset sehingga *clustering* sangat berguna karena dapat mengarah pada penemuan kelompok yang sebelumnya tidak dikenal dalam data. Perbedaan dan kesamaan objek data dinilai berdasarkan nilai atribut yang menggambarkan objek data biasanya melibatkan pengukuran jarak, metode pengelompokan yang berbeda dapat menghasilkan pengelompokan yang berbeda pada suatu dataset. *Clustering* sebagai alat penambangan telah diimplemetasikan diberbagai bidang, terdapat beberapa metode *clustering* dasar antara lain metode partisi, metode hirarki, metode berbasis density, metode berbasis grid.

Metode Partisi adalah salah satu metode *clustering* yang efektif untuk dataset yang berukuran kecil sampai dengan menengah (Wilks, 2011). Partisi data dilakukan secara otomatis oleh algoritma *clustering*.dengan gambaran terdapat seperangkat objek *n*, membangun *k* partisi data, di mana setiap partisi mewakili sebuah *cluster* dan $k < n$, membagi data menjadi kelompok *k* sehingga setiap kelompok harus mengandung setidaknya satu objek. Dengan kata lain, metode partisi melakukan partisi satu tingkat pada set data. Metode partisi dasar biasanya mengadopsi pemisahan *cluster* eksklusif. Setiap objek harus benar-benar milik satu kelompok. Kriteria umum dari *cluster* yang baik adalah bahwa objek dalam satu *cluster* yang sama "dekat" atau terkait satu sama lain mempunyai homogenitas tinggi, sedangkan objek dalam kelompok yang berbeda “berjauhan” atau sangat berbeda mempunyai nilai heterogenitas yang tinggi. Metode partisi tradisional bisa diperluas untuk pengelompokan ruang bagian yang akan berguna ketika ada banyak atribut dan datanya jarang (Han dkk., 2012).

Tipe *cluster* berdasarkan keanggotaan titik data ke grup yang diidentifikasi antara lain (Simovici dan Djeraba, 2014) :

a. *Exclusive or strict partitioning clusters*

   Cluster partisi eksklusif, setiap objek data milik satu *cluster* eksklusif.

b. *Overlapping clusters*

   Grup *cluster* tidak eksklusif, dan setiap objek data mungkin milik lebih dari satu *cluster*. Ini juga dikenal sebagai *cluster* multi-view. Sebagai contoh, seorang pelanggan suatu perusahaan dapat dikelompokkan dalam satu kelompok pelanggan yang untung tinggi dan satu kelompok pelanggan volume tinggi pada saat yang bersamaan.

c. *Hierarchical clusters*

   Setiap *cluster* anak dapat digabung untuk membentuk kluster induk. Sebagai contoh, *cluster* pelanggan yang paling menguntungkan dapat dibagi lagi menjadi *cluster* pelanggan jangka panjang dan *cluster* dengan pelanggan baru dengan pembelian bernilai tinggi

d. *Fuzzy or probabilistic clusters*

Setiap titik data milik semua kelompok *cluster* dengan berbagai tingkat keanggotaan dari 0 hingga 1. Misalnya, dalam dataset dengan kluster A, B, C, dan D, suatu titik data dapat dikaitkan dengan semua *cluster* dengan derajat A50.5, B50.1, C50.4, dan D50. *Fuzzy clustering* mengaitkan kemungkinan keanggotaan ke semua cluster.

e. *Prototype-based clustering:*

Dalam pengelompokan berbasis prototipe, setiap *cluster* diwakili oleh objek data pusat, juga disebut *prototype*. *Prototype* dari masing-masing *cluster* biasanya merupakan pusat dari *cluster*, oleh karena itu, *clustering* ini juga disebut *clustering centroid* atau *clustering* berbasis pusat.

f. *Density clustering*

Cluster juga dapat didefinisikan sebagai wilayah padat di mana objek data terkonsentrasi dikelilingi oleh area dengan kepadatan rendah di mana objek data jarang. Setiap area padat dapat ditetapkan sebagai *cluster* dan area dengan kepadatan rendah dapat dibuang. Dalam bentuk pengelompokan ini tidak semua objek data berkerumun karena objek noise tidak dialihkan ke *cluster* apa pun.

g. *Hierarchical clustering*

Merupakan proses di mana hierarki *cluster* dibuat berdasarkan jarak antara titik data. Keluaran dari *cluster* ini adalah dendrogram: diagram pohon yang menunjukkan berbagai kelompok pada titik presisi yang ditentukan oleh pengguna

h. *Model-based clustering:*

Model-based clustering mendapatkan dasarnya dari statistik dan model distribusi probabilitas; teknik ini juga disebut pengelompokan berbasis distribusi. *Cluster* dapat dianggap sebagai pengelompokan yang memiliki titik data milik distribusi probabilitas yang sama.

2.2.4. Algoritma *K-Means*

K-Means Clustering merupakan metode *clustering* partisi berbasis *prototype* tempat dataset dibagi menjadi *k-cluster*, pada metode ini pengguna

menentukan jumlah *cluster (k)* yang perlu dikelompokkan dalam dataset, Tujuan dari *K-Means clustering* adalah untuk menemukan titik data *prototype* untuk setiap *cluster*; semua titik data kemudian ditugaskan ke *prototype* terdekat, yang kemudian membentuk sebuah *cluster*. *Prototype* disebut sebagai *centroid* atau pusat *cluster*. Pusat *cluster* dapat menjadi rata-rata dari semua objek data dalam *cluster*, atau objek data yang paling terwakili. *K-Means clustering* menciptakan partisi (*k*) dalam ruang *n*-dimensional, di mana *n* adalah jumlah atribut dalam *dataset* yang diberikan. Untuk mempartisi dataset, ukuran kedekatan harus didefinisikan. Ukuran yang paling umum digunakan untuk atribut numerik adalah jarak *Euclidean*. (Simovici dan Djeraba, 2014). Berikut adalah tahapan tahapan proses *K-Means clustering*:

1. Inisialisasi pusat data *(centroid)*

Jumlah *cluster k* harus ditentukan oleh pengguna dalam hal ini akan dioptimasi dengan metode *Elbow*. Langkah pertama dalam algoritma *K-Means* setelah penentuan jumlah *cluster* adalah memulai penentuan *centroid*. Dalam penelitian ini akan dilakukan percobaan membandingkan penggunaan nilai *centroid* awal berdasarkan rumus nilai tengah (*median*), rumus nilai rata-trata (*mean*) dan secara acak (*random*). Berikut adalah rumus *centroid* awal berdasarkan nilai rata-rata (*mean*) (Sarkar dan Rashid, 2016)

$$\mu_i = \frac{x_1 + x_2 + x_3 + \cdots + x_j}{j} \tag{2.2}$$

dengan μ_i adalah nilai rata-rata *cluster*, x_j nilai objek dalam *cluster* dan j jumlah objek dalam *cluster*, sedangkan rumus nilai *median* dengan jumlah data ganjil sebagai berikut (Sarkar dan Rashid, 2016):

$$ME = X_{\left(\frac{n+1}{2}\right)} \tag{2.3}$$

dengan ME adalah nilai *median*, *n* adalah jumlah data, dan X adalah nilai objek namun jika jumlah data genap menggunakan rumus :

$$ME = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \quad (2.4)$$

Nilai *median* juga dapat didapat dengan menjumlahkan nilai terendah dengan nilai nilai tertinggi dibagi 2 dengan rumus sebagai berikut (Sopian, 2017):

$$X^* = \frac{nilai\ terendah + nilai\ tertinggi}{2} \quad (2.5)$$

2. Pengalokasian data

Setiap data atau objek dilokasikan ke *cluster* terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak antar kedua obyek tersebut. Jarak paling dekat antara satu data dengan satu *cluster* tertentu menentukan suatu data masuk ke dalam *cluster* yang mana dalam konteks ini yang "terdekat" dihitung dengan ukuran kedekatan. Jarak *Euclidean* adalah ukuran kedekatan yang paling umum, meskipun ukuran lain seperti ukuran *Manhattan* dan koefisien *Jaccard* dapat digunakan. Jarak *Euclidean* (*d*) antara dua titik data *x (x1, x2, ..., xn)* dan *c (c1, c2, ..., cn)* dengan *n* atribut. Berikut adalah rumus Jarak *Euclidean* (Draisma dkk., 2016)

$$d(x,c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots . + (x_n - c_n)^2} \quad (2.6)$$

dengan d adalah jarak titik *x* dan c, x_n adalah data kriteria dan c_n adalah *centroid* pada *cluster* ke *n*.

3. Hitung *Centroid* Baru

Untuk setiap *cluster*, setelah tahap pengalokasian data iterasi pertama selesai kemudian dihitung *centroid* baru untuk pengalokasian data iterasi selanjutnya. *Centroid* baru ini adalah titik data paling representatif dari semua titik data di *cluster* diperoleh dari rata-rata *cluster*. Secara matematis, langkah ini dapat dinyatakan sebagai meminimalkan jumlah kesalahan kuadrat (*SSE*) dari semua titik data dalam sebuah *cluster* ke pusat masa dari *cluster*. Tujuan keseluruhan dari

langkah ini adalah untuk meminimalkan *SSE* dari masing-masing kelompok. Berikut adalah rumus penetapan *centroid* baru

$$\mu_i = \frac{1}{j_i} \sum_{x \epsilon C_i} X \tag{2.7}$$

dengan μ_i adalah pusat untuk gugus baru ke-*i*, C_i adalah *cluster* ke-*i*, j_i adalah banyak titik data dalam *cluster i* dan *X* adalah objek data spesifik.($x_{1,}x_{2,\ldots\ldots,}x_n$). *Centroid* dengan *SSE* minimal untuk *cluster* yang diberikan *i* adalah rata-rata baru dari *cluster*. *SSE* dihitung menggunakan rumus sebagai berikut (Purnima dan Arvind, 2014):

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \tag{2.8}$$

dengan *SSE* adalah jumlah kuadrat error untuk semua objek dalam himpunan, x_j adalah vektor objek data ($x_{1,}x_{2,\ldots\ldots,}x_n$), C_i adalah *cluster* ke i, μ_i adalah *centroid cluster* ke *i* dan k adalah jumlah *cluster* yang terbentuk.

4. Penghentian

Langkah menghitung *centroid* baru, dan langkah untuk menugaskan titik data ke *centroid* baru diulangi sampai tidak ada lagi perubahan dalam penugasan titik data yang terjadi. Dengan kata lain, tidak ada perubahan signifikan pada *centroid* yang dicatat. *Centroid* akhir dideklarasikan sebagai *prototype cluster* dan digunakan untuk menggambarkan keseluruhan model pengelompokan.

2.2.5. Metode *Elbow*

Metode *Elbow* adalah metode yang terlihat pada persentase varians dijelaskan sebagai fungsi dari jumlah *cluster*. Metode ini ada atas gagasan bahwa seseorang harus memilih *k* jumlah *cluster* sehingga menambahkan *cluster* lain tidak memberi pemodelan data yang jauh lebih baik. Metode ini digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan melihat

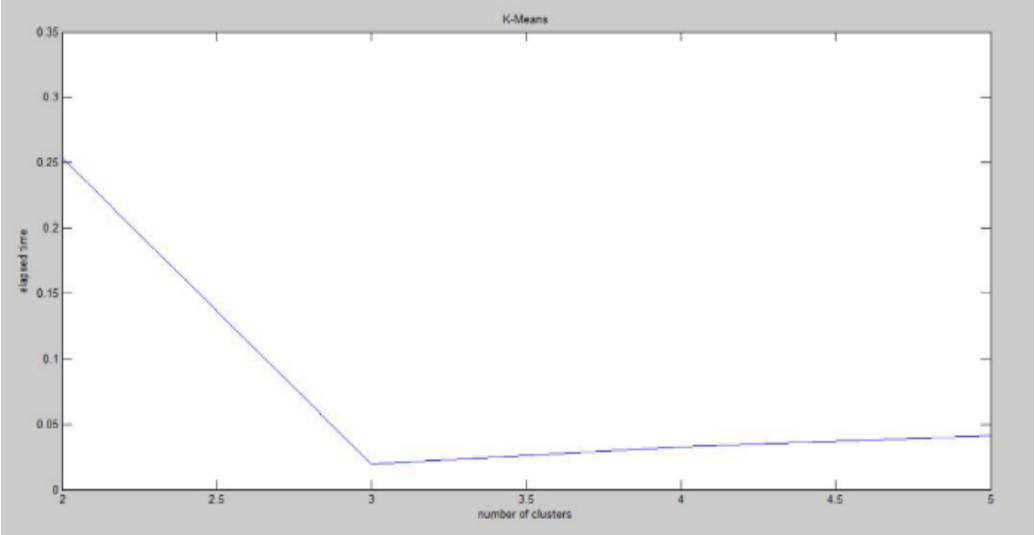
persentase hasil perbandingan antara jumlah *cluster* yang membentuk siku pada suatu titik.

Metode *Elbow* dengan gagasan memilih nilai *cluster*, kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik. Dan selain itu persentase perhitungan yang dihasilkan menjadi pembanding antara jumlah *cluster* yang ditambah. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut siku pada grafik atau nilainya mengalami penurunan paling drastis maka nilai *cluster* tersebut yang terbaik untuk digunakan dalam proses pengelompokan.

Berikut adalah algoritma metode *Elbow* untuk menentukan nilai (*k*) terbaik pada *K-means clustering* (Purnima dan Arvind, 2014) :

1. Inisialisasi nilai *k*
2. Mulai
3. Naikkan nilai *k*
4. Hitung hasil *sum of square error* (*SSE*) setiap nilai *k*
5. Amati hasil *sum of square error* (*SSE*) setiap nilai *k* yang turun secara drastis
6. Tetap kan nilai *k* optimal
7. selesai

Berikut adalah ilustrasi hasil dari Metode *Elbow* untuk menentukan nilai jumlah *cluster* yang terbaik untuk digunakan dalam proses *K-Means clustering*.



Gambar 2. 1 Grafik SSE terhadap jumlah *cluster* (Purnima dan Arvind, 2014)

Dari ilustrasi gambar 2.1 menunjukkan bahwa hasil dari Metode *Elbow* untuk menentukan jumlah *cluster* terbaik yang digunakan dalam proses *clustering* adalah 3 *cluster* yang ditunjukkan pada garis yang membentuk sudut siku.

2.2.6. Validasi *Cluster*

Cluster yang baik adalah *cluster* yang memiliki kesamaan (homogenitas) yang tinggi antar anggota dalam suatu *cluster* dan memiliki perbedaan (heterogenitas) yang tinggi antara *cluster* yang satu dengan *cluster* yang lainnya. Heterogenitas *cluster* dapat digunakan untuk validasi *cluster* dapat dihitung dengan menggunakan rumus (Liu dkk., 2010; Halkidi dkk., 2002) :

$$RS = \frac{SS_T - SS_W}{SS_T} \tag{2.9}$$

dengan

$$SS_T = \sum_{j=1}^{d} \left(\sum_{k=1}^{N} \left(x_{kj} - \mu_j \right)^2 \right) \tag{2.10}$$

dan

$$SS_w = \sum_{j=1}^{d} \left(\sum_{k=1}^{Ni} (x_{kj} - \mu_j)^2 \right) \quad (2.11)$$

dengan RS adalah ukuran heterogenitas *cluster*, SS_T adalah jumlah kuadrat total, SS_w adalah jumlah kuadrat dalam *cluster*, d adalah jumlah dimensi data, x_{kj} objek data ke k dalam *cluster* dimensi d, μ_j mean atau rata-rata data dimensi ke *j, Ni* adalah jumlah data dalam *cluster* ke *i.*