SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

Optimasi Cluster pada Algoritma K-Means

Mira Orisa [1)]

[1),2),3)]*Teknik Informatika, Institut Teknologi Nasional Malang*
Jl. Sigura-gura 2 Malang
Email : mira.orisa@lecturer.itn.ac.id

Abstrak. *Metode evaluasi yang digunakan adalah metode-metode internal. Metode internal melakukan evaluasi dengan melihat seberapa jauh jarak antar cluster dan seberapa padat cluster-cluster tersebut. Pengklasterisasian data dimodelkan menggunakan algoritma K-Means. Algoritma K-Means memiliki kelemahan dalam menentukan centroid awal. Centroid awal ditentukan secara random/acak untuk sejumlah k cluster yang dipilih. Sehingga keluaran yang dihasilkan bergantung pada pemilihan centroid awal tersebut. Algoritma K-Means harus dijalankan berulang kali untuk mendapatkan hasil cluster yang optimal. Evaluasi cluster untuk menemukan jumlah cluster terbaik pada algoritma K-means dapat ditentukan dengan metode internal seperti metode Elbow, Davies Bouldin Index, dan Silhouette Index. Metode Elbow merupakan Teknik evaluasi internal yang mengukur evaluasi cluster dengan Sum of Square Error(SSE). Metode Davies Bouldin Index mengukur evaluasi cluster dengan Sum of Square Within Cluster (SSW) dan Sum of Square Between Cluster (SSW). Sedangkan metode silhouette index menggunakan perhitungan nilai koefisien. Hasil optimasi cluster menggunakan metode elbow yaitu jumlah cluster optimal adalah 3 dengan titik elbow berada di k=3. Sedangkan untuk hasil optimasi untuk metode davies bouldin index dan silhouette index yaitu jumlah cluster optimal adalah 2 dengan jumlah nilai DBI terendah ada di k = 2 yaitu sebesar 0.3228986726354396 . SI yang mendekati 1 adalah di k=2 sebesar 0,894.*

Katakunci: *K-Means, elbow, davies bouldien index, silhouette index.*

1. Pendahuluan

Dewasa ini perkembangan data semakin heterogen dan kompleks dengan volume yang terus meningkat. Tentu saja pengolahan *big data* akan mengalami kesulitan baik dalam membaca data ataupun dalam menemukan pola-pola dan relasi-relasi data jika dilakukan secara manual. Dalam *data mining* dapat dilakukan pengelompokan data yaitu salah satunya menggunakan metode *clustering* menggunakan algoritma *K-Means*. Klasterisasi menggunakan algoritma *K-Means* telah banyak digunakan seperti klasterisasi minat siswa pada pelajaran matematika[1], segmentasi konsumen[2], untuk klasterisasi data rekam medis pasien[3], untuk mengolah data *web usage mining*[4], klasterisasi tingkat penjualan paket data telkomsel[5], untuk pengelompokkan produktivitas tanaman pada di Jawa Tengah[6], untuk menganalisis penjualan sebuah toko fashion hijab[7], untuk strategi pemasaran produk industry kreatif untuk mengelompokkan data UMKM[8], untuk mengelompokkan data kinerja dosen universitas islam bandung[9], untuk klasterisasi data obat pada rumah sakit asri[10].

Algoritma *K-Means* mengelompokkan data berdasarkan kedekatan/kemiripan data sehingga data-data yang memiliki karakteristik yang sama akan dimasukkan kedalam *cluster* yang sama. Kedekatan data tersebut bisa diukur dari seberapa dekat jarak antar data dan jarak data dengan *centroid* data. Prinsip kerja algoritma *K-Means* yaitu menentukan jumlah k *cluster* terlebih dahulu kemudian menentukan titik *centroid* data secara random/acak. Setelah itu mengalokasikan data ke *cluster* terdekat dan proses tersebut akan diulang hingga menemukan *centroid* yang stabil. Keluaran algoritma *K-Means* sangat bergantung pada penentuan jumlah *cluster* dan pemilihan *centroid* awal yang ditentukan secara random/acak. Permasalahan yang ada pada algoritma *K-Means* adalah menghasilkan *centroid* akhir yang tidak benar-benar menjadi pusat *cluster* yang sesungguhnya. Dalam prakteknya algoritma ini harus dijalankan berkali-kali dengan *centroid* awal yang berbeda-beda untuk mendapatkan *centroid* akhir yang dianggap paling baik[11].

Metode evaluasi *cluster* dapat menyelesaikan masalah tersebut. Metode evaluasi *cluster* Seperti metode *Elbow, Davies Bouldin Index* Dan *Silhouette Index* merupakan metode internal yang dapat membantu untuk mendapatkan klasterisasi ideal pada algoritma *K-Means*.

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

Prinsip dasar algoritma *K-Means* adalah[12]:

1. Menentukan *centroid*

   Centroid adalah pusat *cluster*. Tiap-tiap cluster memiliki *centroid*. Penentuan *centroid* awal dilakukan secara random.

2. Menentukan anggota *cluster*

   Melakukan pengelompokan data/objek ke tiap-tiap *cluster*. Pengelompokkan tersebut dihitung berdasarkan jarak antar data/objek ke *centroid* masing-masing *cluster* yang sudah ditentukan sejak awal. Data/objek yang paling dekat dengan *centroid* maka akan dikelompokkan menjadi satu *cluster* dengan *centroid* tersebut.

Adapun Langkah-langkah kerja algoritma *K-Means* berdasarkan prinsip diatas adalah[13]:

1. Menentukan jumlah k *cluster*
2. Menentukan *centroid* tiap *cluster* secara random/acak
3. Mengalokasikan semua data/objek terdekat dengan *centroid* terdekat dengan data/objek. Pada umumnya penentuan jarak terdekat antar data/objek dengan *centroid* menggunakan rumus *Euclidean distance* sebagai berikut:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots .. + (x_{ip} - x_{jp})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

   Dimana:
   d(I,j) = jarak data ke I ke pusat *cluster* j
   x_{ki} = data ke i pada atribut data ke k
   x_{kj} = titik pusat ke j pada atribut ke k

4. Ulangi Langkah tersebut hingga mendapatkan *cluster* yang stabil yaitu dimana nilai *centroid* baru tidak berubah-ubah lagi atau sama dengan nilai *centroid* lama.

Pemilihan jumlah k *cluster* optimal dapat menggunakan metode *Elbow, Davies Bouldin Index* atau *Silhouette Index*. Metode *Elbow* dapat menentukan jumlah k *cluster* optimal dengan membandingkan persentase antar *cluster* yang membentuk siku pada titik tertentu. Informasi hasil dari metode *Elbow* ditunjukkan dengan grafik. Untuk mendapatkan perbandingan tersebut menggunakan nilai *Sum of Square Error* sebagai berikut[13]:

$$SSE = \sum_{K=1}^{K} \sum_{x_{ie}\, S_K} \|x_i - C_K\|_2^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Dimana x_i dan C_K adalah:

x_i = nilai atribut dari data ke i

C_K = nilai atribut titik pusat *cluster* ke i

Davies bouldin index (DBI) menentukan jumlah k *cluster* optimal berdasarkan nilai kohesi dan separasi data/objek. Kohesi adalah jumlah kedekatan data/objek terhadap *centroid* nya dalam satu *cluster*. Separasi adalah jarak antar *centroid* dari *cluster*. Untuk mencari matrik kohesi menggunakan *Sum of Square Within Cluster (SSW)* sebagai berikut[13]:

$$SSW_I = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

Dimana m_i , c_i, $d(x_j, c_i)$ adalah:

m_i = jumlah data dalam *cluster* ke i

c_i = *centroid cluster* ke i

$d(x_j, c_i)$ = jarak tiap data/objek ke *centroid* yang dihitung dengan metode jarak *Euclidean*

Sedangkan mengetahui separai data/objek menggunakan *Sum of Between Cluster (SSB)* sebagai berikut[13]:

$$SSB_{I,J} = D(C_I, C_J) \quad \text{.....(4)}$$

Kemudian akan dicari nilai rasio untuk mengetahui nilai perbandingan antara *cluster* ke i dan *cluster* ke j, dihitung dengan persamaan[13]:

$$R_{ij} = \frac{SSW_I + SSW_J}{SSB_{i,j}} \quad \text{.....(5)}$$

Nilai perbandingan rasio digunakan untuk mencari nilai DBI, sebagai berikut[13]:

$$DBI = \frac{1}{K} \sum_{k=1}^{k} \quad max_{i \neq j}(R_{i,j}) \quad \text{.....(6)}$$

Dimana K adalah jumlah *cluster* yang digunakan.

Silhouette Index menampilkan ukuran kedekatan setiap titik data/objek dalam satu *cluster* dengan titik-titik data/objek *cluster* tetangga. Pada metode *Silhouette Index* menggunakan koefisien a dan b Koefisien a adalah rata-rata jarak antar data/objek dalam satu *cluster*. Dan koefisien b adalah rata-rata jarak terkecil terhadap semua data/objek dari *cluster* tetangga. Bentuk persamaan matematinya[13]:

$$a_i^j = \frac{1}{m_{j-1}} \sum_{r=1 \; r \neq 1}^{m_j} \quad d(x_i^j, x_r^j), \quad i = 1,2, \ldots . . m_j \quad \text{.....(7)}$$

dimana $d(x_i^j, x_r^j)$ = jarak data/objek ke i dengan data/objek ke r dalam satu *cluster* j

m_j = jumlah data/objek dalam *cluster* ke j

$$b_i^j = min_{n=1,\ldots k \; n \neq j} \left\{ \frac{1}{m_n} \sum_{r=1 \; r \neq 1}^{m_n} \quad d(x_i^j, x_r^n) \right\}, \quad i = 1,2, \ldots . . m_n \quad \text{.....(8)}$$

Setelah nilai koefisien a dan b diperoleh maka akan digunakan untuk mendapatkan nilai SI dengan persamaan sebagai berikut[13]:

SI data ke i

$$SI_i^j = \frac{b_i^j - a_i^j}{max\left\{a_i^j, b_i^j\right\}} \quad \text{.....(9)}$$

SI *cluster* digunakan untuk menghitung rata-rata nilai SI semua data/objek dalam satu *cluster*

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \quad SI_i^j \quad \text{.....(10)}$$

SI Global

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

$$SI = \frac{1}{k}\sum_{j=1}^{k} \quad SI_j$$
…….. (11)

Dimana k= semua *cluster*

Nilai SI yang mendekati 1 menandakan bahwa sebuah data semakin tepat berada dalam *cluster* tersebut, jika nilai SI negatif $a_i > b_i$ menandakan bahwa data tidak tepat berada dalam *cluster* tersebut, atau lebih dekat dengan *cluster* lain. Sedangkan jika nilai SI bernilai 0 berarti bahwa data/objek berada di perbatasan antara dua *cluster*[13].

Data yang digunakan untuk klasterisasi adalah data statistik covid-19 kapupaten/kota di jawa timur pertanggal 29 Juni 2022 hingga 5 Juli 2022(https://infocovid19.jatimprov.go.id). Data tersebut akan diproses melalui tahapan *Knowledge Discovery in Database* (KDD). Adapun tahapan *Knowledge Discovery in Database* (KDD)seperti[14]:

1. *Selection*
   data pertama kali akan dilakukan proses seleksi untuk menentukan *variable* yang akan digunakan.
2. *Preprocessing*
   Setalah data diseleksi maka akan dilakukan proses *cleaning* data untuk menghilangkan *missing value,* atau data yang tidak konsisten dan tidak relevansi. Dan dilakukan integrasi data untuk atribut-atribut yang mengidentifikasi entitas yang unik.
3. *Transformation*
   Pada tahap ini dapat disesuaikan dengan algoritma yang digunakan dalam *Data Mining*.
4. *Data Mining*
   Disinilah data akan ditambang untuk mendapatkan informasi atau model tertentu menggunakan metode-metode dalam *Data Mining.*
5. *Evaluation*
   Evaluasi adalah tahap penilaian untuk hasil dari proses *Data Mining* apakah sudah memenuhi target yang ditentukan atau belum.
6. *Knowledge*
   Tahap terakhir dari KDD ini adalah agar pengetahuan yang diperoleh dapat memberikan manfaat kepada penggunanya.

2. Pembahasan

Data yang digunakan untuk proses klasterisasi menggunakan algoritma *K-Means* merupakan hasil proses *selection, preprocessing* dari data statistik covid-19 kapupaten/kota di jawa timur pertanggal 29 Juni 2022 hingga 5 Juli 2022. *Variable* yang digunakan untuk klasterisasi adalah kabupaten/kota, jumlah terkonfirmasi, jumlah sembuh, dan jumlah yang meninggal. Data ditunjukkan pada Tabel.1.

Tabel 1. Data statistik covid-19 jawa timur(per 29 Juni 2022-5 Juni 2022)

No	Kabupaten/Kota	Terkonfirmasi	Sembuh	Meninggal
1	Kab.Malang	25887	24825	1054
2	Kab.Banyuwangi	19421	17565	1848
3	Kab Nganjuk	16528	15640	885
4	Kab Sidoarjo	45213	44151	1019
5	Kab Jember	21332	19842	1487
6	Kota Madiun	10142	9590	548
7	Kab Jombang	16642	14993	1637
8	Kab Lumajang	11280	10236	1040
9	Kab Gresik	20923	20164	748
10	Kan Situbondo	8889	7993	895
11	Kota Malang	28925	27648	1247
12	Kab Sumenep	6391	6106	284

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

No	Kabupaten/Kota	Terkonfirmasi	Sembuh	Meninggal
13	Kab Bangkalan	7595	6834	755
14	Mojokerto	11158	10920	232
15	Kab Trenggalek	9677	8625	1051
16	Kab Lamongan	8986	8542	444
17	Kab Ponorogo	14677	13240	1437
18	Kota Surabaya	117623	114565	2947
19	Kota Blitar	7741	7471	269
20	Kab Kediri	21185	19890	1294
21	Kota Kediri	5682	5280	401
22	Kota Probolinggo	5674	5291	383
23	Kota Pasuruan	4870	4610	258
24	Kab Tuban	9129	8187	938
25	Kab Tulungagung	10437	10143	286
26	Kab Sampang	3576	3408	165
27	Kota Batu	4940	4647	287
28	Kab Pamengkasan	3180	2952	228
29	Kota Mojokerto	5213	4971	241
30	Kab Probplingo	8809	8267	542
31	Kab Pasuruan	11850	11114	735
32	Kab Bojonegoro	9628	9000	627
33	Kab Ngawi	10480	9537	943
34	Kab Bondowoso	8356	7573	782
35	Kab Pacitan	9734	9412	318
36	Kab Blitar	12973	11333	1637
37	Kab Magetan	12857	11802	1054
38	Kab Madiun	11007	10295	712

Penilaian optimasi *cluster* untuk *dataset* pada Tabel 1 menggunakan metode *Elbow* dengan menggunakan persamaan (2). Untuk mendapatkan perbandingan persentase antar *cluster* yang membentuk siku.Hasil perhitungan *Sum of Square Error* untuk semua k uji ditunjukkan pada Tabel 2.

Tabel 2. Hasil *SSE* semua k uji

No	Jumlah k *cluster*	Nilai *SSE*
1	2	4758058165,18
2	3	1641257905,72
3	4	691332925,92

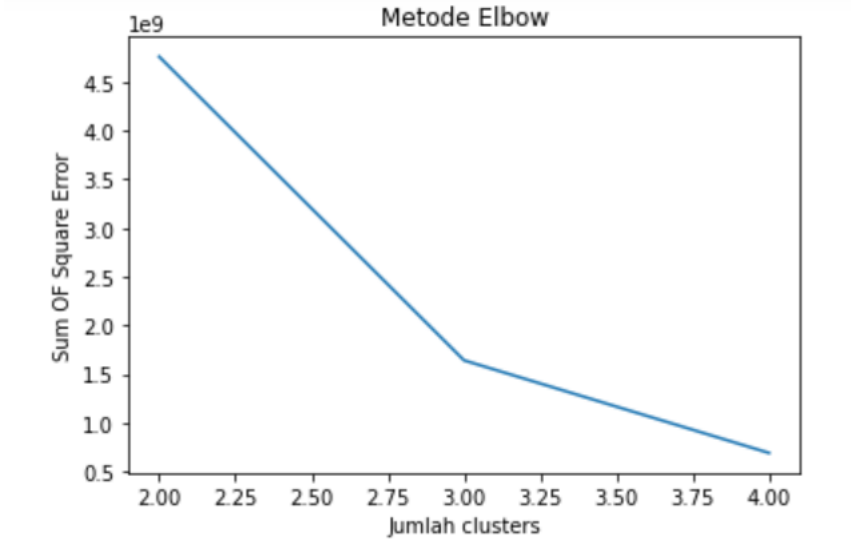
Untuk menentukan pesentase antar *cluster* yang membentuk siku pada suatu titik dengan menghitung selisih tiap nilai *SSE* untuk semua k uji. Dilihat dari hasil selisih nilai *SSE* k uji pada Tabel 3, selisih terbesar ada di titik k = 3

Tabel 3. Hasil selisih *SSE* semua k uji

No	Jumlah k *cluster*	Nilai *SSE*	Selisih
1	2	4758058165,18	0
2	3	1641257905,72	3116800259
3	4	691332925,92	949924979,8

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

Oleh sebab itu maka *cluster* optimal adalah k=3 dan grafik siku ditunjukkan pada Gambar 1.



Gambar 1. Identifikasi titik *Elbow Berdasarkan SSE*[penulis]

Sedangkan optimasi *cluster* menggunakan *Davies Bouldin Index* dari sejumlah k uji ditunjukkan pada Tabel 4. Terlihat bahwa nilai *DBI* terendah ada di k = 2 yaitu sebesar 0.3228986726354396. sehingga menurut perhitungan metode *Daviel Bouldin Index* yang merupakan *cluster* optimal adalah dua *cluster*.

Tabel 4. Hasil selisih *SSE* semua k uji

No	Jumlah k *cluster*	Nilai DBI
1	2	0.054399178595 98228
2	3	0.380875769578 7312
3	4	0.260625020947 4576

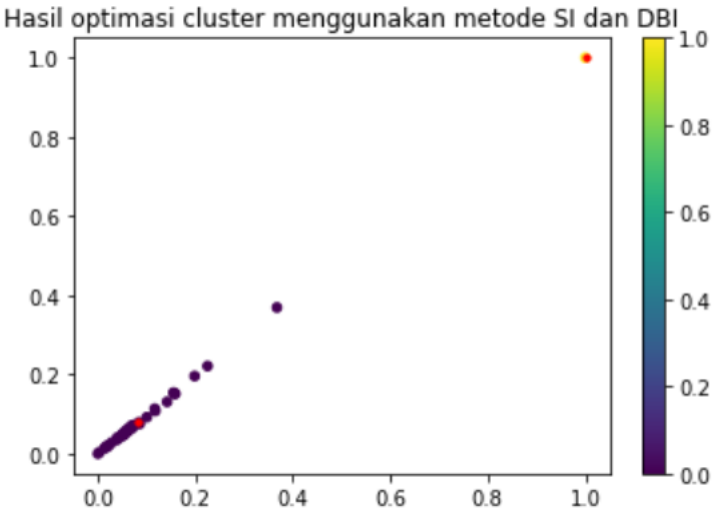
Untuk hasil perhitungan pada metode *silhouette index* diperoleh nilai *rata-rata dari SI cluster* dari semua k uji ditunjukkan pada Tabel.5. Terlihat bahwa hasil perhitungan *SI* yang mendekati 1 adalah di k=2 sebesar 0,894.

Tabel 5. Hasil *Silhouette Index* pada semua k uji

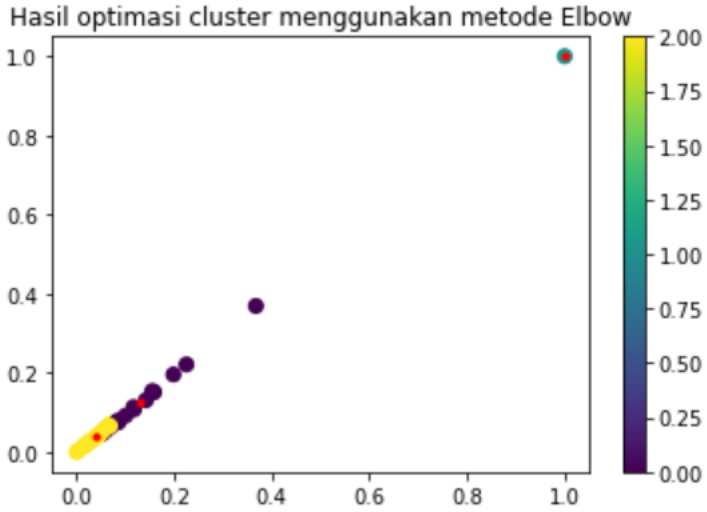
No	Jumlah k *cluster*	Nilai *SI*
1	2	0,894
2	3	0,669
3	4	0,642

Hasil Optimasi *cluster* pada algoritma *K-Means* menggunakan metode *Elbow* divisualisasikan pada Gambar 2. Sedangkan hasil Optimasi *cluster* pada algoritma *K-Means* menggunakan *Silhouette Index* dan *Davies Bouldin Index* divisualisasikan pada Gambar 3.

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022



Gambar 2. Hasil optimasi *cluster* dengan metode *SI* dan *DBI*[penulis]



Gambar 3. Hasil optimasi *cluster* dengan metode *Elbow*[penulis]

3. Simpulan

Ada beberapa hal yang perlu diperhatikan dalam mengolah data untuk data mining. Tidak hanya bagaimana proses dari algoritma yang digunakan tetapi juga harus memperhatikan bagaimana dataitu di seleksi dan dibersihkan dari data yang tidak konsisten. Data yang dijadikan inputan pada *algoritma K-Means* harus sudah melalui tahapan *selection, clensing data, dan transformation.* Berdasarkan hasil evaluasi untuk menentukan jumlah *cluster* yang optimal diperoleh jumlah *cluster* yang paling baik menurut perhitungan metode E*lbow* adalah tiga *cluster* dengan nilai k= 3 memiliki selisih paling besar diantara semua k uji yaitu 3116800259. Sedangkan menurut perhitungan metode *Davies Bouldin Index* dan *Silhouette Index* menghasilkan bahwa jumlah *cluster* yang paling baik berjumlah dua *cluster*. Terlihat bahwa di k=2 untuk memiliki nilai *Davies Bouldin Index* terendah ada di k = 2 yaitu sebesar 0.3228986726354396. Dan hasil perhitungan *Silhouette Index* yang mendekati 1 adalah di k=2 sebesar 0,894.

Daftar Pustaka

[1]. Arofah. Siti Nur & Fitri Marisa,2018.Penerapan Data Mining Untuk Mengetahui Minat Siswa Pada Pelajaran Matematika Menggunakan Metode *K-Means Clustering*. Journal of Information Technology and Computer Science.Vol.3,No.2.e-ISSN:2541-6448

SENIATI 2022
ISSN 2085-4218
ITN Malang, 13 Juli 2022

[2]. Auliasari.Karina & Mariza Kertaningtyas,2019.Penerapan Algoritma *K-Means* untuk Segmentasi Konsumen Menggunakan R.Jurnal Teknologi & Manajemen Informatika.Vol.5,No.1.
[3]. Ali.Amir,2019.Klasterisasi Data Rekam Medis Pasien Menggunakan Metode *K-Means Clustering* di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo.Jurnal MATRIK.Vol.19,No.1,Hal 186-195.e-ISSN:2476-9843
[4]. Orisa,Mira & Michael Ardita,2020.*Web Usage Mining* Menggunakan Algoritma *Clastering K-Means*.Jurnal Teknologi Informasi dan Terapan (J-TIT).Vol.8,No.1.ISSN:2580-2291
[5]. Handoko,suhandio, fauziah,Endah Tri esti Handayani,2020.Implementasi Data Mining Menentukan Tingkat Penjualan Paket data Telkomsel Menggunakan Metode *K-Means Clustering*.Jurnal Ilmiah Teknologi dan Rekayasa.Vol.25,No.1.
[6]. Wijayanto,Sena & M Yoka Fathoni,2021.Pengelompokkan Produktivitas Tanaman Padi di Jawa Tengah Menggunakan Metode *Clustering K-Means*.Jurnal JUPITER.Vol.13,No.2,Hal 212-219
[7]. Normah,Siti Nurajizah, Arinda Salbinda,2021.Penerapan Data Mining Metode *K-Means Clustering* untuk Analisa Penjualan pada Toko fashion Hijab Banten.Jurnal Teknik Komputer AMIK BSI.Vol.7,No.2.E-ISSN:2550-0120
[8]. Herdiana,Oding,Shanti Maulana,Eryan Ahmad Firdaus,2021.Strategi Pemasaran Produk industry Kreatif Menggunakan Algoritma *K-Means Clustering* Berbasis *Particle swarm optimization*.Jurnal Nuansa Informatika.Vol.15,No.2.e-ISSN:2614-5405.
[9]. Rachmasari,Shavira Siti,abdul Kudus,2021.Perbandingan Penerapan Algoritma *K-Means* dan *Fuzzy C-Means* untuk Mengelompokkan data Kinerja Dosen Universitas Islam Bandung.Prosiding Statistika.Vol.7,No.2.
[10]. Nugroho,Muhamad Rizki,Iwansyah edo Hendrawan,Puwantoro,2022.Penerapan algoritma *K-Means* untuk Klasterisasi Data Obat pada Rumah sakit ASRI.Jurnal Nuansa Informatika.Vol.16,No.1.E-ISSN:2614-5405.
[11]. Suyanto,2019.Data Mining untuk klasterisasi dan Klasifikasi Data.Penerbit:informatika.Bandung.ISBN:978-602-6232-97-7.
[12]. Purnama.Bedy,2019.Pengantar Machine Laerning Konsep danPraktikum dengan Contoh Latihan Berbasis R dan Python.penerbit:Informatika Bandung.ISBN:978-623-7131-19-9.
[13]. Jollyta,Deny ,Muhammad Siddik,herman Mawengkang,Syahril efendi,2021.Teknik evaluasi *Cluster* solusi Menggunakan python dan Rapidminer.Penerbit:Deepublish.yogyakarta.ISBN:978-623-02-2303-7.
[14]. Febianto,Nugroho Irawan & Nico Dias Palasara,2019.Analisis *Clustering K-Means* pada data Informasi Kemiskinan di Jawa Barat tahun 2018.Jurnal SISFOKOM.Vol.8,No.2.E-ISSN:2581-0588.